Fall 2023

Framing Automation Trust: How Initial Information About Automated Driving Systems Influences Swift Trust in Automation and Trust Repair for Human Automation Collaboration

Scott Anthony Mishler
Old Dominion University, smishler17@gmail.com

Recommended Citation

Mishler, Scott A.. "Framing Automation Trust: How Initial Information About Automated Driving Systems Influences Swift Trust in Automation and Trust Repair for Human Automation Collaboration" (2023). Doctor of Philosophy (PhD), Dissertation, Psychology, Old Dominion University, DOI: 10.25777/qtn2-k597
https://digitalcommons.odu.edu/psychology_etds/417

FRAMING AUTOMATION TRUST: HOW INITIAL INFORMATION ABOUT AUTOMATED DRIVING SYSTEMS INFLUENCES SWIFT TRUST IN AUTOMATION AND TRUST REPAIR FOR HUMAN AUTOMATION COLLABORATION

by

Scott Anthony Mishler
B.S. December 2016, Purdue University
M.S. May 2019, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

PSYCHOLOGY

OLD DOMINION UNIVERSITY
December 2023

Approved by:

Jing Chen (Co-Director)

Mark Scerbo (Co-Director)

Jeremiah Still (Member)

Hong Yang (Member)

ABSTRACT

FRAMING AUTOMATION TRUST: HOW INITIAL INFORMATION ABOUT AUTOMATED DRIVING SYSTEMS INFLUENCES SWIFT TRUST IN AUTOMATION AND TRUST REPAIR FOR HUMAN AUTOMATION COLLABORATION

Scott Anthony Mishler
Old Dominion University, 2023
Director: Dr. Jing Chen

The study examines how trust in automation is influenced by initial framing of information before interaction and how later active calibration methods can further influence trust repair or dampening after an automation error in a three-experiment study. As more human drivers begin to use automated driving systems (ADSs) for the first time, their initial understanding of the system can influence their trust leading to a miscalibration of trust. Prior studies have investigated how trust develops through interactions with an automated system, but few have looked at integrating swift trust and framing to calibrate trust before interaction and investigate further active calibration methods after an error. We conducted three experiments using multiple drives with an ADS to test manipulations of user's initial trust calibration, how resilient that trust manipulation would be to an automation error, and if the trust could be repaired or further dampened after the error. Three initial framing methods were employed before the drives: Positive/Promotion, Control, and Negative/Dampening. The second experiment implemented an error during one of the drives and the third experiment implemented a positive, control, or negative active trust calibration strategy after the error. Positive/Promotion framing did indeed show an increase in trust for the first experiment and that increased trust was resilient in drives after an error in the second experiment. However, the third experiment showed mixed results and was unable to demonstrate an effect of active trust calibration after the error. Overall, the study showed that framing information is impactful on trust for driver's initial interactions with an ADS, certain active calibration methods might not be effective depending on the individual, and designers and researchers should be careful to consider these effects to avoid endorsing overtrust or undertrust.

This dissertation is dedicated to my parents, Ty and Christy Mishler, for always encouraging me to learn and inspiring me to achieve greatness. I am lucky to be able to follow their example of dedication and determination. I am forever grateful for their unconditional love and support throughout my life.

ACKNOWLEDGMENTS

It is hard to count the number of people who have helped me on my academic path to this point, but I would like to express my thanks to everyone I have met along the way that has contributed, either to my knowledge and research or to just keeping me sane along the way. I want to express my appreciation for my advisor, Dr. Jing Chen, as she has been invaluable not only throughout the dissertation process, but through my entire graduate school experience. Her guidance helped me to reach this point and I could not have done it without her. A special thanks to Dr. Robert W. Proctor, without whom I would not have discovered the field of Human Factors. My undergraduate experience in his lab and his Human Factors course shaped my knowledge and fueled my passion for research. I appreciate my committee, Dr. Mark Scerbo, Dr. Jeremiah Still, and Dr. Hong Yang for all the numerous contributions to improve the study and document. An additional thanks to Dr. Yusuke Yamani who served on my quals committee for his efforts and insightful questions. I would also like to thank my fellow lab member Katie Garcia for her aid and the undergraduate research assistants that helped with data collection and analysis: Melody Jing, Sharazad Ali, Kevin Hernandez, James Xu, Manuela Urbanske Donnelly, Kyle Canady, Kayla MacPherson, and Joey Flores.

NOMENCLATURE

ADOS	Automated Driving Opinion Survey
DDT	Dynamic Driving Task
LKA	Lane Keeping Assistance
ADS	Automated Driving System
AV	Automated Vehicle
ACC	Adaptive Cruise Control

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Automation is an ever-increasing aspect of human life, and it is important to ensure that humans and automation work cooperatively to promote safety and efficiency. However, neither humans nor automation are perfect and therefore it is important to understand the interactions between them to ensure optimal workload distribution. Collaborative teaming between humans and automation has been a large push in recent literature with more improvements and better automation coming out (Endsley, 2017; Hancock, 2017). A commonly referenced statistic about traffic accidents reports that 94% of accidents are due to the human driver, often reported as due to human error (Read et al., 2021; Singh, 2018). The majority of these driver-related reasons are usually recognition, decision, or performance errors. Advanced automotive assistance systems (ADAS) have been shown to help for such cases (Benson et al., 2018; Cicchino, 2017).

However, implementation of these systems has shown sometimes either unwillingness to use the systems that result in loss of benefits (de Visser et al., 2020; Parasuraman & Manzey, 2010; Sanchez et al., 2014) or overreliance on the system that leads to issues if automation errors occur (Dunn et al., 2021; Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997). Ensuring proper trust calibration between the human-automation team, much like human-human teams (Madhavan & Wiegmann, 2007), is crucial for safety (de Visser et al., 2020; Parasuraman & Riley, 1997). For some new and unknown systems, swift trust, which is defined as quickly formed trust in situations with little prior interaction that require fast cooperation due to heightened risk or stakes, needs to be introduced to calibrate trust to the capabilities of the system (Capiola et al., 2020; Haring et al., 2021). Creating trust quickly can help create teams that work together more efficiently for new groups (Kroeger et al., 2021; Meyerson et al., 1996).

In human-machine collaboration, calibrating trust for different automation systems can require trust repairs due to error or dampening if trust is too high (de Visser et al., 2018, 2020). In ADAS, many systems can be new to drivers or are constantly updated and might not be fully implemented or understood (DeGuzman & Donmez, 2021b). Integrating trust repair and dampening techniques should improve trust calibration, particularly in situations where assistance systems might be novel to the user.

1.1 Autonomous Vehicles

Autonomous vehicles are currently a hot topic, yet their future is uncertain. Certain experts have predicted that we would have fully autonomous vehicles by this point and some still put that prediction far off into the future. In the meantime, we are stuck in an intermediary period where there are automated driving system (ADS) features that help the human driver navigate the roadway and watch out for hazardous situations. Features like Adaptive Cruise Control (ACC) and Lane Keeping Assistance (LKA) help drivers maintain a consistent speed with the flow of traffic and stay in the proper lane without drifting. These ADS features allow the human driver to offload some of the driving task and focus on monitoring other potential hazards. As the technology improves, driving will become increasingly automated. On the transition to highly automated driving (HAD), many drivers will be on the forefront of the technology and will have to learn how it works along the way.

According to the SAE levels of driving automation (SAE, 2021) there are six levels of automation from 0 – no automation – to 5 – full automation – with the current commercially available ADSs ranging around 2 and HADs that companies are working toward at level 4. Level 2 automation only requires that the vehicle performs the steering and acceleration, and the human driver monitors and is ready to take control. At Level 2 automation system continuously assists the driver by integrating both acceleration and steering in certain conditions using technologies like highway pilot. The human driver is still required to be fully attentive and engaged throughout the entire driving task. Level 3 gives more responsibility to the vehicle, and it monitors the roadway environment, but the human is still required to take over. At Level 3, the automation now takes over the complete object and event detection and response (OEDR), meaning that the human only needs to monitor the automation's performance, but is not required to maintain attention on the roadway environment unless the vehicle requests their action. At Level 4, the automation completes the full driving task but there are limitations on where this can be performed, leaving the driver the option for override but not the requirement within those confines.

Automation errors or limitations can occur at Levels 2 to 3 as system-limit or system-malfunction (DeGuzman et al., 2020; Mishler & Chen, 2023). System-limit errors occur when the system is no longer able to perform the task because the boundaries of its capabilities have been reached, like when the road conditions change to one outside the operational definition or the weather conditions prevent proper functioning. In contrast, the system-malfunction errors occur when the system should be functioning properly in its operational conditions, such as when the system does not see a vehicle slowing down and react by slowing down accordingly. Compensating for these automation errors requires some cooperation between the automation system and the human driver. Much discussion has surrounded the idea of who is responsible in cases of crashes or incidents during these levels of driving, but that is mostly outside the scope of this study. However, it is important to point out that at Levels 2 and 3 there is a *shared responsibility* of the driving task wherein both the human and the automation must take actions that can change the trajectory of the vehicle (Bennett et al., 2020; SAE, 2021).

Most car companies shy away from explicitly labeling their technology along the levels put forth by the SAE, leading to some difficulty fully classifying current systems and creating room for confusion. Additionally, manufacturers have made promises of higher advance automated driving that still have not come to fruition, leaving question about when these future levels might actually be feasible and if they will be able to fully perform with only the driver as the fallback or without human need (Ford Motor Company, 2016). This resistance to official labeling is likely to continue into the future as the technology improves and makes the jump to levels 3 and 4, during which human becomes further removed from the process, but automation is required to be much more reliable and robust. Current automation classification schema might change and adapt as the technology grows and many questions about automation levels are already in question because of the lack of human-automation teamwork and coordination involved and the focus on what “man does better versus machine does better” (Dekker & Woods, 2002; Jamieson & Skraaning, 2018; Kaber, 2018).

Periodic updates and improvements to automated vehicles are likely needed to improve the automation capability – along with vehicle system function and cybersecurity – over time, whether through over-the-air updates or larger updates while in the workshop. Vehicle software updates, like our phones and computers, could introduce critical changes that adapt the way drivers interact with the system. Even small improvements over time to various driving assistance systems can result in larger changes to the driving behavior and capabilities of the system. The novelty of the update can change the way the individuals perceive, use, understand, and interact with the system. The current study focused on system updates as such.

1.2 Trust in Automation

Trust in automation literature started in the realm of human-human trust, applying principles of how humans interact and build trust with each other to situations of humans interacting with automation (Muir, 1987; Muir & Moray, 1996). Many similarities in the level of trust between human-human pairs and human-automation pairs exist but can differ depending on the initial level of acceptance, anthropomorphism or how humanlike the automated agent is, the reliability of the agent and many other factors. Trust is a psychological state in which one party accepts a certain level of uncertainty of the other party with the expectation of positive results of behavior or subsequent interactions from that party (Rousseau et al., 1998). Applied specifically to human-automation trust, Lee and See (2004) define trust as the belief that an agent will help the user achieve a task or goal when that goal has some level of uncertainty and vulnerability. In both cases, the uncertainty and vulnerability are present because the individual must rely on an external entity that is outside of their control and must believe that the other is both capable and willing to do whatever was agreed upon. Often in human-human interactions, trust can be more uncertain due to the other human having individual desires and goals that might be separate than the individual. However, in human-automation pairs, the automation is typically programmed with the intent of explicitly helping the individual and therefore the issue of trust often comes from the potential of its capability and not the willingness or motivation to help. Although the automation is only there to assist, humans are often prone to the automation bias, expecting much more from automation than their human counterparts, leading to unrealistically high standards for automation (Dzindolet et al., 2003).

Trust changes and develops based on the behavior of the collaborative agent. However, determinants of that trust can vary depending on the nature of that agent. Parasuraman and

Manzey (2010) assert that trust develops differently in humans than in technology because of a common positivity bias towards new technologies. Dzindolet and colleagues (2003) posit that people are potentially unrealistically optimistic about new technologies regarding its capabilities. However, this positivity may not be resilient as mentioned earlier where the automation bias can make humans more critical toward mistakes made by automation (Dzindolet et al., 2003; Goodyear et al., 2017). Additionally, the common positivity toward new technologies does not always hold across the board. Some technologies such as AI-created art show more negative reactions compared to human created art (Hong & Curran, 2019). Trust can develop and increase over time, though errors can lead to trust decreases as commonly seen with technology (Madhavan & Wiegmann, 2007). Humans are also more prone to blaming technology for errors due to the fundamental attribution error (Jones & Harris, 1967)and shifting responsibility for errors while also being unwilling to give automation credit for positive results (Madhavan & Wiegmann, 2007). This is further supported from findings of algorithm aversion where participants did not depend on a forecasting aid after an error but preferred to follow a human input even when the errors were more severe (Dietvorst et al., 2015; Glikson & Woolley, 2020).

Prior literature on human-automation trust has focused on the idea that humans use automation as a tool, and it is either used or not. However, automation's role is shifting from that of only a tool to that of a teammate (Demir et al., 2021). As AI and robots become more capable, they could be seen as collaborators. In such situations, humans can depend on AI to autonomously complete a task while the human works on something else or the human-automation team can work in tandem to accomplish a task. Errors in work are almost guaranteed, by both humans and automation. That is why cooperative teamwork exists to benefit from each other's strengths and make up for individual weaknesses. One member can notice the error of another and help correct it or provide input to help prevent the error again in the future. Cooperative research for automated vehicles has explored communication strategies to avoid misunderstandings that might lead to errors through the automation communicating lane change intent (Kauffmann et al., 2018). Alternatively, automation should not be designed to supplant the human driver, at least for this stage of AVs. The best way to design for AVs would not be something that fully replaces the tasks and capabilities of the human driver, because that would push the human out of the loop (Endsley, 2017; Endsley & Kiris, 1995).

One of the most influential factors influencing trust for human-automation is the reliability of the system (Chen et al., 2021, 2022; Hoff & Bashir, 2015). De Visser and Parasuraman (2011) manipulated both static reliability and adaptive reliability for human-robot teaming of monitoring robotic vehicles and showed significantly higher trust for aids with higher reliability than those with lower reliability. Chen and colleagues (2021) manipulated the reliability of a phishing detection aid to either 60% or 90% reliability and found that trust was significantly higher for the 90% reliable aid. Additionally, they found that with higher reliability, participants trusted the aid more regardless of the error type compared to lower reliability where trust calibration varied depending on if the error was a miss or false alarm. False alarms are more prevalent in less reliable automation, so trust is calibrated down to account for the noticeable errors. But misses for unreliable automation might go unnoticed so trust is never decreased. Seong and Bisantz (2008) showed that reliability information and feedback can significantly contribute to accurate trust calibration by informing users when it is appropriate to remember the aid’s imperfections or ignore the aid when necessary. Showing if an aid is unreliable lets users decide when to use it and highlighting its high reliability still reminds them it is not perfect. Errors demonstrate the reliability of the automation to the operator; however, some errors might result in small trust breaks whereas others are seen as major trust violations (de Visser et al., 2018; Hoff & Bashir, 2015). For errors in automated driving, some studies have shown that takeover requests lower trust (Hergeth et al., 2015; Mishler, 2019; Mishler & Chen, 2023), whereas others have shown that a takeover might not be considered an error depending on how it is presented and thus does not influence trust (Gold et al., 2015; Körber et al., 2018).

The amount of time that the individual spends with the automation can also affect their trust level (Hoff & Bashir, 2015). Trust is built over time and the experiences that the individual has with the automation can influence their current level of trust calibration. However, this is not always the case (Bailey & Scerbo, 2007). Sometimes, operators might not be able to detect the reliability of the automation or notice errors, especially at high levels of automation where the discrepancy between perfect and near perfect automation is very small (Bailey & Scerbo, 2007). Glickson and Woolley (2020) review the current work on human trust in AI in specific regard to the differences between trust in AI and other forms of technology, with a focus on the specific form of AI representation (robot, virtual, embedded). As machine intelligence gets better, that leads to more complex capabilities requiring a higher user understanding to fully grasp the innerworkings of the system (Hancock, et al., 2011). Trust is often derived from a user’s perceived capabilities of the system rather than its true capabilities. An additional complication arises due to the non-deterministic nature of AI meaning that the decision-making process is hard to make transparent and decisions are often hard to predict (Danks & London, 2017).

1.3 Trust Calibration

Trust calibration is the concept through which the level of trust in the automation by the human is adjusted to properly match the capabilities of the automation (Lee & See, 2004; Muir, 1987). Trust calibration is critical for human-automation collaboration because improper calibration can result in either overreliance on the system if the human is too trusting or underreliance if they are not trusting enough. Overreliance on automation is problematic for human-automation teams because it leads to misuse of the system.

One can find many examples in the literature of instances where an operator overtrusted and misused an automated system resulting in catastrophic errors. The 2018 Uber incident where an ADS (Levels 3-5 of automation) hit and killed a pedestrian crossing the road was one of the watershed cases for the real-world consequences of automated driving systems (National Transportation Safety Board, 2018). The general conclusion of the NTSB report was that the ADS was never able to accurately identify the pedestrian even though it was tracking the object; however, if the human driver would have been paying attention, they would have likely had enough time to intervene and avoid the pedestrian. Though the system was being tested at higher levels where the human driver is not required to pay attention, there are situations where the lower level ADASs will fail, and the ADS does not detect it. Some individuals even misuse automation further by overriding safety checks of ADASs to keep the automation going past the operating specifications, like the product called the “autopilot buddy” for Tesla vehicles sold to trick the automation into thinking that the driver’s hands were still on the wheel (Hawkins, 2018). Clearly some individuals have too high of trust in systems that are not yet worthy of such level of trust.

Conversely, automated driving systems could go unused due to underreliance, resulting in disuse. Disuse of highly reliable automation systems can cause a loss of safety that the automation would have otherwise provided through its ability to detect threats and potentially act quicker than the human (SAE, 2021). In other words, the added safety benefits provided by a reliable system might not be realized by drivers who disuse the system. Additionally, if they distrust the system, they may not heed alerts or understand certain features of the automation enough to perform a task where automation is required. With some technologies, people are forced to trust the system because they must use it. In situations like these, disuse can manifest in a resistance to listening to system recommendations or warnings, or purposely sabotaging the system to get around requirements. This can be seen in minor ways like drivers modifying their seatbelts so that the automated sensors do not constantly warn them the seatbelt is unbuckled. The consequences from a lack of proper trust can be monumental without correct calibration.

Trust can be calibrated *through* interacting with the system. Interaction with the system and time spent are among the best ways for the human operator to understand how the system functions and develop a proper level of trust (Chen et al., 2018, 2022; Hergeth et al., 2017; Lee, & See, 2004). However, individuals do not always have sufficient time spent interacting with the system or may stop interacting with a system before they have appropriately calibrated their trust. This is especially true for new and innovative systems, as well as for new updates on known systems. Properly designed features of automation systems such as communication style, system appearance, and transparency can lead to proper trust calibration (de Visser et al., 2012; Dzindolet et al., 2002; Lee, 2008). Transparency is when a system provides information about how it is working or its decision-making process to the user to keep the user in the loop. Improving the transparency of the system helps demonstrate how the system is working and how well it is performing, which has been shown to improve trust and trust calibration, especially in systems where high trust is needed (Hoff & Bashir, 2015; Dzindolet et al., 2002). However, transparency information is often lacking for human-machine systems which can leave the user out of theloop regarding what the system is doing or how the system is going to perform in certain circumstances. A system can convey uncertainty or show performance directly, but many

ADSs have too broad of an operating range or do not have enough testing data to know specifics about its own operational reliability, nor is giving explicit reliability rates always useful (de Visser et al., 2020; Du et al., 2020; Katrakazas et al., 2015; Kunze et al., 2019). Additionally, the way transparency information is communicated to the user can influence their understanding and ability to use the given information.

1.4 Framing Effect

The same information can be presented in different ways, or framed, which affects how decisions are made based on the information (Tversky & Kahneman, 1981). In classic framing-effect experiments, individuals are more risk-seeking when a negative frame is presented compared to a positive frame of a question (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). The studies often frame one question in a positive light by highlighting the savings aspect, such as the number of lives or money that might be saved, in contrast to the negative or loss frame, highlighting the number of lives or money that will be lost. By using proper phrasing and highlighting the positive or negative aspects involved with certain decisions, it is possible to influence individuals' decisions. Individuals do not always think rationally and the way a situation is represented to them can influence their decision or attention to certain aspects of a task.

Some prior research has focused on the effects of attribute and goal framing on decision making. Researchers looked at how displaying an automated aid's actual accuracy or inaccuracy percentages compared to highlighting the general goal of maximizing hits and correct rejections or minimizing misses and false alarms (Lacson et al., 2005). In other words, they wanted to know if specific percentages help more than general goals. They found that positive framing, highlighting the accuracy or maximizing hits and correct rejections influenced participants to better utilize the aid. However, Lacson and colleagues also found that showing both positive and negative forms of information lead to increased compliance rates compared to uninformed groups. This finding shows that more information can sometimes make users simply rely on the automation without questioning it, especially when risk information is hard to understand (Brust-Renck et al., 2013; McLaughlin & Mayhorn, 2014). Research has shown that summary risk information can be helpful for users to establish a mental framework, and safety framing of the summary information influences users to pick less risky options (Chen et al., 2015).

Information can be conveyed in different ways and sometimes simply displaying specific reliability information can have unintended consequences, especially when the risk information is hard to understand (Brust-Renck et al., 2013; Lacson et al., 2005). Depending on the transparency information the user is told and how it is communicated, it can influence their trust and performance during subsequent interactions with the system. However, the way the system information is described to users can greatly influence their understanding of how the system works. Singer and Jenness (2020) told participants about a driving assistance system, with the system being called “AutonoDrive” for half of the participants, and for the other half it was called “DriveAssist”. Participants who received AutonoDrive materials thought the system was much more capable than it really was and were more prone to risky behaviors during system use.

Similarly, with updates to a system giving it new capabilities, the brief descriptions of these updates to the user are important, as seen from previously discussed framing studies with minor wording changes. System updates are ubiquitous among technology because features are constantly being added and improved, or bugs are getting fixed. Some short patch update notes are relatively common and can quickly inform individuals about what has changed, as with small security warnings (Chen et al., 2015, 2018; McLaughlin & Mayhorn, 2014). Communicating system capabilities to users through quick messages about the aid can influence users’ perception

(Bass et al., 2013; Mishler et al., 2019; Pak et al., 2012). ADSs do not require individuals to be trained before use and as previously discussed, even when individuals have access to training documentation, usually very little time is spent reading it.

1.5 Trust Promotion, Dampening, and Repair

Sometimes it can be helpful to adjust a user's trust in the system to help with trust calibration *before* their interaction with the system. Even though users are constantly calibrating their trust in automation through interaction with the system, sometimes starting too trusting or too resistant is undesired. Additionally, trust adjustments before the interaction can be useful as an error might be rare or the system has learned from a previous error. Calibrating trust before interaction helps users set reasonable expectations for future interactions. It is important for proper trust calibration to set proper expectations for an upcoming task. This is the case for both the user, who has little experience with the new system update, and the system that has is not yet proven its competence in a new update.

As previously discussed in the framing section, positive or negative framing of messages can impact the way a user understands information and consequently makes a decision. Specifically for conveying reliability information, displaying either too much information or hard to understand information can simply lead to compliance instead of proper trust calibration (Brust-Renck et al., 2013; Lacson et al., 2005). Positive framing of information can lead users to see more benefits, thus highlighting the system's usefulness. Whereas negative framing can lead to better understanding of the risks, leading to more cautious behavior. Setting expectations for how an automation is going to perform can be accomplished through this initial framing before interaction.

Positive framing can result in *trust promotion*, highlighting the benefits of the automation and improving the individual's trust and confidence in the system. Under-reliance on a system, or disuse, can stem from uneasiness about how capable the system is and reluctance to trust new technology, especially when there is limited experience (Parasuraman & Riley, 1997). As will be discussed later, quickly inciting trust in a system can be critical to successful cooperation, especially for new teams (Haring et al., 2021; Kroeger et al., 2021; Meyerson et al., 1996).

In contrast, negative framing can result in *trust dampening*, highlighting the risks or potential points of failure of an automation can set realistic understandings of the automation's capabilities and reliability. In some cases, individuals are prone to misuse of the automation wherein their overreliance on the automation leads them to neglect proper monitoring behavior (de Visser et al., 2020; Lee, & See, 2004). Trust dampening is intended to reduce the amount of overtrust a user has in the automation to dissuade them from overreliance thus insuring a more appropriate level of trust calibration (de Visser et al., 2020). Trust dampening is important for calibrating trust to avoid danger of omission errors through missing critical signal due to lowered monitoring or being too slow to respond due to overreliance and lack of situation awareness. Overtrust can be an unhealthy adaptive strategy that humans adapt when they are in situations of complacency or when burdened with a secondary task (Hancock et al., 2011; Hoff & Bashir, 2015; Parasuraman & Manzey, 2010). In Level 3 autonomous driving, the human driver is not intended to actively monitor the driving environment but must be ready to take over if the situation requires (SAE, 2021). The human driver is likely engaged in a secondary, non-driving related task (NDRT) and thus is highly prone to relying on the automation to safely perform the driving task. Even for lower levels of automation where the human is supposed to be in control of the driving task, human drivers are prone to engaging in secondary NDRTs, resulting in accidents due to distracted driving (Overton et al., 2015). Both trust promotion and dampening strategies can act as preventative measures against improper trust calibration before first interacting with a system, combatting misuse or disuse. However, if these strategies are misapplied, they could end up doing more damage.

During interaction with the system, it is important to ensure trust stays properly calibrated *after* an error. A system error is a violation of trust that should result in a decrease of trust (Chancey et al., 2017; Hoff & Bashir, 2015; Mishler & Chen, 2023). However, errors are typically not an if, but a when. Even a professional athlete or writer with decades of experiences can make a critical error. *Trust repair* is the idea that one member of a team tries to prevent or recover from trust loss after an error by increasing trustworthiness (de Visser et al., 2018; Kim et al., 2006). The concept is familiar in human-human trust situations where one party makes an error, and they implement certain strategies to correct their behavior in the eyes of the other party so that the task operation can continue. One method of cooperative interaction for human-automation teams is the regulation of relationship equity through repair strategies after an error (de Visser et al., 2020).

Trust repair strategies are useful for recalibrating a human's trust after an automation error. If the human does not know why the error occurred, they might believe that it may occur again soon, therefore decreasing trust which could result in an incorrect trust calibration level. De Visser and colleagues (2018) list a series of repair strategies, a few key strategies are to apologize (Kim et al., 2004, 2006), explain (Dzindolet et al., 2003), and deny (Kim et al., 2004, 2006). Using the apologize strategy, the system would make a simple apology for the fault. The explain strategy would describe to the user the reason for the error they just experienced. For the deny strategy, the system would tell the user that it did not make an error. An apology accepts the blame whereas the denial tries to place the blame elsewhere. The explain strategy implements some form of system transparency, allowing the user to see what might have gone wrong in the process and allow them to look out for the potential of this issue in the future. After a violation of trust, an effort to repair this trust can ensure a healthy working relationship between the team. Trust repair is especially helpful in situations where there is an initial bias against trust by one party and certain errors might confirm their lack of trust, even if those errors are rare (Baker et al., 2018; Marinaccio et al., 2015). Said differently, if an individual is biased to disuse a system, a small error might confirm this bias. This is especially pertinent for human-automation teams where humans have unrealistically high standards for automation, which can lead to more negative reactions to automation than to a human teammate (Dzindolet et al., 2003; Goodyear et al., 2017). A well-timed trust repair strategy could encourage the team to continue to work together but would allow the user to be more watchful of the potential for an error in the future and to be more prepared for it, while still benefiting from the mutual teamwork in the meantime.

The process of trust development for automation, from introduction of the system to how the human responds to and recovers from errors, is highly important for understanding how users will interact with the system. De Visser and colleagues (2018) provide an illustration of the transactional model of trust repair and demonstrate how relationship acts can be beneficial (repair, politeness, positive interaction) or costly (error, damage, time loss, inefficiency). These acts are adjusted through a relationship regulation act such as a repair act or a dampening act (i.e., a repair act counteracts a costly act, and a dampening act counteracts a beneficial act). They also demonstrate possible trust recovery trajectories over time with small declines in trust after mini trust breaks, a large drop after a major violation then baseline recovery with no intervention, fast recovery after one repair effort, and delayed recovery for another repair effort.

These depictions set up the theoretical underpinnings behind trust development and calibration. However, this framework only covers repair and dampening *after* interaction and does not consider trust dampening or trust promotion strategies *before* interaction and interaction with newly formed teams. Additionally, these models need more research and validation through further demonstration of human-automation teams. This is where a gap in the research lies and further contributions to the field of human-automation interaction can be made.

My prior work (Mishler, 2019; Mishler & Chen, 2023) demonstrated that trust could be developed and calibrated over seven short six-minute drives using an ADAS. For the condition with no failures on any of the drives, there was a positive increasing linear trend of trust. However, for the conditions with a takeover request and a system-malfunction (system did not avoid construction event in the lane and did not warn the driver) on the fourth drive, there was a significant drop in trust. On subsequent drives after the takeover request and system-malfunction, there was a steady positive improvement of trust similar to the trend of baseline recovery proposed by de Visser and colleagues (2018). Active trust repair strategies in the field of human-robot interaction have shown positive results for the ability to recover trust (Alarcon et al., 2020; Esterwood & Robert, 2021). Although some research in the human-human team trust field has found less importance for active trust building strategies that actively test trust early in the relationship (Kroeger et al., 2021). However, this active trust was dependent on individuals testing each other and not using strategies like apologize or giving transparency information. Other strategies for managing driver’s trust in AVs have even attempted to have the vehicle observe the driver’s behavior and compare that to the current driving context, then using different communication styles to repair or dampen trust for the driver (Azevedo-Sa et al., 2020).

In practical applications, perfect trust calibration or even good calibration is hard to quantify. It is hard to measure reliability of ADS/AVs in the real world (Kalra & Paddock, 2016) and would take months or years of testing to establish a relatively accurate level of reliability. During that hypothetical testing time, new iterations, developments, and changes to the software and performance would alter the results. Endsley (1995) commented on a similar topic regarding the proper calibration of situation awareness an individual would need to have. She concluded that the amount of situation awareness one would need is fully reliant on how much probability of error one is willing to accept (Endsley, 1995). She added that perhaps instead of an overall requirement of situation awareness, it should require specific levels of situation awareness for certain components at definite points in time. The same can be said for trust in automation, which is almost inherent in the common definition: a willingness to use a system based on uncertainty and vulnerability (Lee & See, 2004). Many situations during driving can be variable in terms of risk and level of ability required, and ADSs have different functionality depending on conditions and operating modes. So, understanding how certain situations influence trust in automation, especially when the automation is new to an individual or has new features can be vitally important.

1.6 Initial Systems Interaction and Experience

When experiencing new technology, a novice user would rather be safe than sorry (Annett & Stanton, 2006). The novice user tends to have a conservative bias when interacting with a new system, meaning that they predict a higher number of errors than what actually occurs. New systems are complex in terms of human-automation interaction because improved performance of the system may be masked by other factors in evaluation testing due to its newness (Endsley, 1995). A new automated system might perform better than its earlier version, but a user might not have enough experience with the system to properly trust it and gain its full advantage. For instance, the individual might not be proficient in using the new system so their understanding of the task and performance might be lowered while they acclimate but could improve after adaption.

When new features of a technology are introduced, it is much the same as that of a human collaborator attempting a new task that they have not yet established competency in. A human learning a new instrument and then giving a performance with others might be expected to make a few mistakes along the way but will correct their mistakes and improve. Similar with an automated system introducing a new feature and requiring a bit of time to work out the bugs. However, small development errors can be especially difficult for a human collaborator to understand if they are unfamiliar with how the task or system works (de Visser et al., 2020; DeGuzman & Donmez, 2021a, 2021b). In certain systems like adaptive cruise control (AAC), drivers have limited knowledge about how the system works even if they own it (DeGuzman & Donmez, 2021a). Even with the introduction of anti-lock braking systems many years ago, many drivers did not understand the technology and had unintended reactions to the system such as taking their foot off the brake because they were unfamiliar with the vibration (Smiley, 2000). A study by Singer and Jenness (2020) found that a majority of participants knew some of the limitations of ADASs such that it does not work in harsh weather or when road lines are faded, but a majority thought that ACC would be able to take action to avoid a collision if a car ahead brakes quickly. As the level of automation increases and new technologies are introduced, the automation system could see improvements to the limitations and would be able to avoid collisions that were not previously possible, as with Level 3 automation. Additionally, individual elements of a system might adjust, change, and improve. A Level 3 automation vehicle might have features and capabilities tied to lower mode levels and individuals might react differently to changes in these isolated elements (Lee, 2018).

First interactions with a system are much different than those after familiarity has been established. With the future integration of Level 3 automation and the capabilities involved, it is important to understand how these new systems are adapted and understood by the human driver. Initial experiences with an automation system or new feature might heavily impact trust especially if there are errors early on as compared to later (Fox & Boehm-Davis, 1998; Lee, & See, 2004). A large portion of the human automation interaction literature focuses on how experts or experienced individuals with training interact with automated systems that have been around for a while such as nuclear power plants or process control systems. Trust is more prone to fluctuation early on and less resilient compared to individuals experienced with a system, especially when reliability varies (Chiou & Lee, 2021; Fox & Boehm-Davis, 1998). Even a subject matter expert using a newly developed system or feature will have different insights and experiences interacting with the system for the first time than an individual who is unfamiliar with the task or process. Commercial pilots are required to undergo many hours of training, must be certified, and are re-evaluated regularly. Thus, they are more experienced in their task than the typical driver of an automobile who has a relatively short and less in-depth training. Whereas a typical road vehicle driver would receive no formal training or evaluation and might not even be fully aware of the functionality of the new system, though new ISO standards for automotive updates could have certain requirements, depending on adoption (International Organization for Standardization, 2023).

Familiarity with a system highly influences the calibration of trust. Initial interactions with a system or feature can have a more pronounced effect on trust, but after long exposure to a system, user trust is more resilient (Hoff & Bashir, 2015). Tenhundfeld et al. (2019) tested the effect of familiarity on trust by manipulating familiarity through providing either verbal information about the autopark feature or giving a demonstration of how to use the autopark feature to participants using the automated parking feature of a Tesla. The demonstration provided a higher level of familiarity than simply giving the participant information. The lack of familiarity resulted in more distrust of the system as shown through a higher rate of intervention – the human taking over control of the vehicle during the parking process – by the participants in the information condition. However, as they gained more familiarity with the system through multiple parking trials, the rate of intervention decreased. The automated parking system still made some slight errors and had to correct itself sometimes, but participants were able to calibrate their trust based on this behavior through experience to reach an appropriate level for this system. However, this experiment did not manipulate the way the information was provided to the individual. Hands-on experience is clearly more useful than general information, but how the information is presented and what type of information is conveyed can have varying effects on trust. Similar results were also found for conveying information regarding the automated aid reliability through description versus experience (Chen et al., 2018; Mishler et al., 2017). They found that descriptive information provided to users about an aid's reliability improved subjective trust in the system, but only feedback on users' performance via experiences improved both subjective and objective trust calibration. Initial descriptive information could be helpful at first, but later interaction helps to verify the level of trust, leading to proper calibration.

Sometimes, users do not act according to expectations even when provided with system limitation information. Victor et al. (2018) trained participants on the limitations of an ADS and found that 28% of participants did not take over control of the vehicle in time to avoid a hazard even though they were informed of the limitations. Semi-structured interviews revealed that the individuals who did not avoid the collisions trusted that the vehicle would handle the situation. Related research has found similar findings, showing that participants were prone to overtrust and complacency. Participants did not switch from “hands-free” to “hold the steering wheel” even when warned at both 60 seconds and 15 seconds in advance and remained hands off for other times when the Tesla autopilot was engaged (Banks et al., 2018). Other research has shown that training focused on de-emphasizing the system’s limitations results in higher trust in the system compared to training that emphasizes the limitations (Körber et al., 2018). In a survey of non-owners and owners of vehicles with ACC and LKA systems, researchers found that knowledge of system capabilities influenced trust for non-owners with little experience, but not for owners with experience in the system (DeGuzman & Donmez, 2021b). DeGuzman and Donmenz conclude that it could be more beneficial to focus on simply making individuals aware that the system is fallible and reinforcing the human driver’s role in use of the systems.

1.7 Swift Trust

Swift trust refers to trust that needs to be developed rapidly among individuals or groups in the absence of extended prior interactions. Or, more formally as summarized from the seminal swift trust article of Meyerson and colleagues (1996), it is an initial expectation that upon first interaction between members of a group, the other party will meet performance expectations even though no prior experience or concrete evidence has led them to this conclusion. The fundamental aspect defining swift trust is that it does not rely on a history of interaction experience to inform trust but must arrive at a quick judgement to trust. The basis for the decision to trust comes from sources other than experience (Meyerson et al., 1996). The importance of swift trust has been seen for human-human teams in organizations where cooperation between colleagues with no interpersonal history is commonplace (Bakker, 2010). Even on a wider scale, in day-to-day interactions with individuals like healthcare providers, public transit drivers, and rideshare drivers, and more require reliance on their expertise without prior history.

Swift trust relies on importing information from various sources based on stereotypes about the agent. Separate from information gathering, which may take additional time, imported trust is quickly acquired, similar to a downloaded dataset or template based on predetermined heuristics (Meyerson et al., 1996; Wildman et al., 2012). Compared to traditional sources of trust where it is built through interaction, the initial trust is assumed and then checked later. The imported information consists of information from other sources like their qualifications, group membership, job/title, and more. The information an individual receives about the other agent could greatly impact their initial trust formation or swift trust (McKnight et al., 1998). Kroeger and colleagues (2020) discuss how the role of the individual agent impacts swift trust. A role is defined as what the agent is required to do in the teamwork process. They found that when roles are clear, trust is augmented, but when roles are blurred or inconsistent, trust is compromised. Greenberg and colleagues (2007) discuss how one's within-role ability inspires swift trust. Individuals are not trusted simply because they fit the role, but because they seem capable of fulfilling the duties of that role given the specific task. Likening this to automated driving, individuals do not simply trust an AV to perform because that is its role but need more context of either ethos of the vehicle's abilities, reputation of the brand, or seeing a brief performance to know the tasks that fit the driving role can be accomplished. Role clarity is not often provided to the user in partial automation driving, with even sales personnel not fully understanding the roles (Endsley, 2017). Human drivers often do not fully understand what they are expected to do for a driving task and what the ADS is responsible for (Degani et al., 1999; Sarter & Woods, 1995; Tesla, 2017).

Swift trust is also supposed to be for semi-temporary groups that have indefinite time together. However, that amount of time is not well defined, and some research has studied swift trust for just initial meetings or over an entire semester for a virtual online course with a professor (Coppola et al., 2004). Swift trust is ephemeral, as the concept no longer applies after a certain amount of time because the two agents are interacting and thus grow to know each other as time continues. The core idea of swift trust is that it is initially assumed until further interactions can either justify the trust or adjust their beliefs as needed (Langfred, 2004); of course, the adjustment of trust and transition into interaction-based trust models would follow this initial swift trust development stage. The common goal of the team necessarily requires initial trust to function but trust calibration through later interaction and more long-term developments come into play (Haring et al., 2021). All of that considered, it is a useful framework to understand initial trust levels in both human-human and human-automation teams as they begin before transitioning to a longer-term model of trust.

Swift trust literature has focused on human-human interactions, but the same basic interactions can originate in the human-automation realm such as the first time interacting with an unknown technology such as an ATM, a new phone app, or an automated parking meter. However, we have not seen much application of swift trust to this human-automation, though there have been some researchers trying to fill this gap in human-robot interaction (Haring et al., 2021). The role of an individual is a contributor to swift trust and understanding another’s role can help inform the other individual of what to expect (Kroeger et al., 2021; Meyerson et al., 1996). When meeting with a medical doctor, one can know what to expect in terms of their credentials and experience based on the role of the doctor. The same can be said for human-automation pairings where a certain type of technology and its role can help the human understand what they can expect. However, further definition of certain roles might be helpful to clarify what type of individual action is required in that role (Coppola et al., 2004; Kroeger et al., 2021). For example, knowing the doctor's specialty or even information about their prior behavior from a friend can help specify the role.

Members of human-automation or human-robot pairings might have even less antecedents of swift trust than traditional human-human pairings due to the nature of automation (Haring et al., 2021). It is more difficult for an automated agent to convey their goals or abilities than for a human, especially during an initial interaction. Humans have more verifiable credentials such as job experience or degrees attained compared to a potential untested or unknown automation. Though branding information associated with automation could play a role in user confidence, it is much more variable and less tangible than something like a person's job experience. Humans have more innate surface level cues (e.g., age, appearance, facial expression, body language) that are easily observable compared to an automation partner, although some automation can be designed to emulate these human cues utilizing anthropomorphism through a humanistic look or voice. However, not all automation needs to or can be anthropomorphized. It is helpful to understand how an automated system can convey information to a human because the automation lacks many of the traditional interaction cues that typical human-human teams have. Communicating initial role information or importing information for human-automation teams may be more complicated than human-human pairs, thus it is necessary to introduce these concepts into the literature and further explore this gap.

The way initial information is conveyed by the automation to the human could vastly change their swift trust and thus affect the later development of trust through extended interaction.

CHAPTER 2

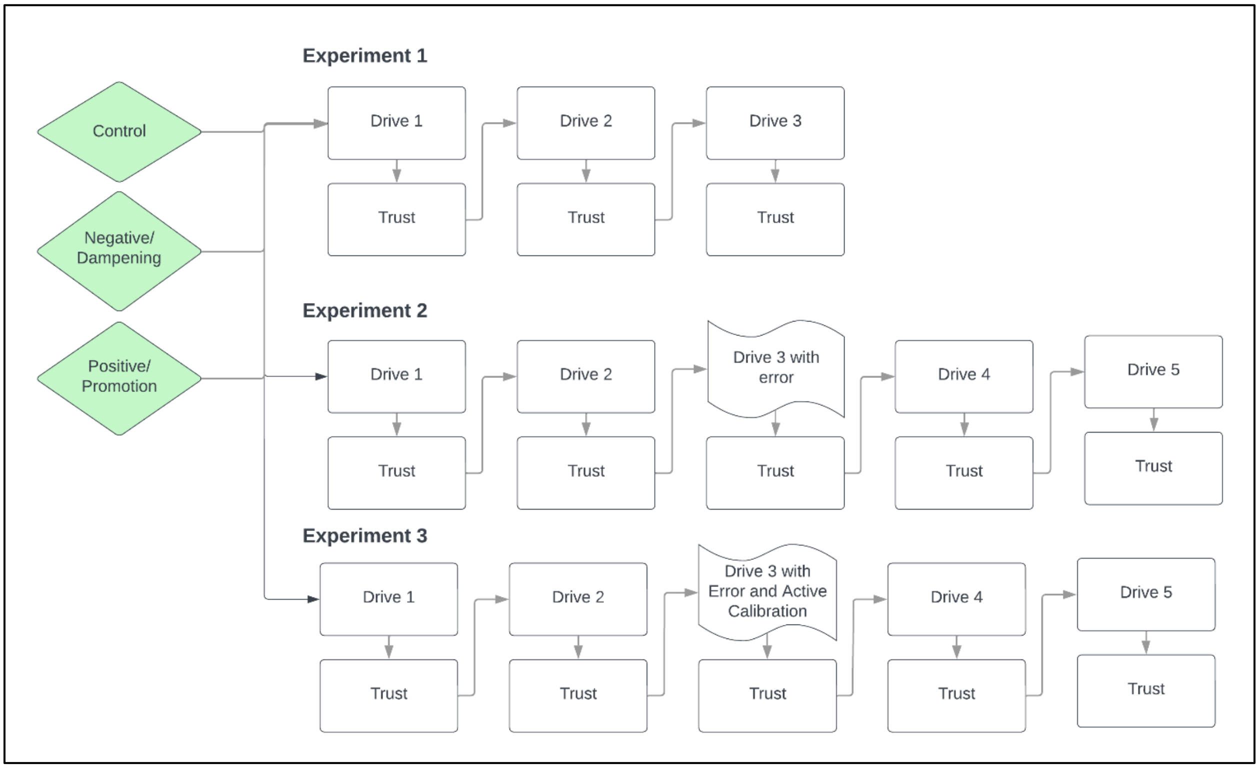
EXPERIMENTS

The goal of these experiments was to understand the initial, swift trust development process of new human-automation teams after an automation update and calibrate trust after automation errors to support proper trust calibration. Considering how critical proper trust in automation is for using an ADS, and how variable some framing and trust designs can be, the first experiment investigated how the presentation of update information can influence human drivers' trust through positive and negative framing. The initial interaction after an automation update between the human and automation is perfect for testing swift trust between human-automation teams because the team is brand new with no prior experience.

This first experiment ensured that the results would properly align with the desired goals as the ideas and applications are in a new domain with potentially uncertain designs. The first experiment demonstrated calibrations of trust over three drives using an ADS after either a positive (trust-promotion), negative (trust-dampening), or neutral control framing of an update. After demonstrating the expected differences due to framing and supporting the hypotheses, Experiment 2 expanded the length to five drives and added an automation error to see how swift trust begins transitioning to learned trust after interaction and how the framing effects their trust in response to an error and the subsequent trust recovery after the error. The third experiment examined ways of actively calibrating trust following an error. Figure 1 shows the overall procedure of each of the three experiments.

Figure 1

The Overall Procedure of Each of the Three Experiments



2.1 Experiment 1

Positive and negative framing can be used to promote or dampen a specific feeling such as trust. By focusing on the positive aspects of a system or choice, the benefits are highlighted which in turn promotes more positive feelings about the object. Whereas damping can highlight the negative or uncertain features of a system which can help users understand the limitations of the system and adjust accordingly.

This study aimed at understanding how the presentation of a new update of an ADS can influence human drivers' trust calibration when the update information was presented as positive or negative framing. Typically, system updates will notify a user that new features are available but does not illustrate the potential downfalls or benefits, leaving a user in the dark about the system's new capabilities regarding trust. The experiment was set up to demonstrate the differences in calibration of trust over three drives using an ADS after either a positive, negative, or neutral framing of an update.

2.1.1 Hypotheses

The first hypothesis was a confirmatory hypothesis that trust would increase over time (increasing trust linearly from drives 1-3; H1). Increasing trust through experience has been shown in numerous driving studies and is a staple of trust research findings (Gold et al., 2015; Hoff & Bashir, 2015; Kraus et al., 2019; Mishler & Chen, 2023). Trust in a system should continue to be built quickly and evenly because the system does not make any errors and swift trust that exists prior to interaction is being confirmed (Hancock et al., 2011; Kroeger et al., 2021; Sanchez et al., 2014).

Hypothesis 2 was that the positive framing of update information would result in greater trust than both control and negative framing; and conversely, that negative framing would result in lower trust than both control and positive framing (H2). This result of framing would be expected from both prior framing research and transparency research showing that information about the system's performance helps users calibrate trust to the capabilities of the system (de Visser et al., 2012; Dzindolet et al., 2002; Tversky & Kahneman, 1981).

Hypothesis 3 predicted that trust calibration would stay consistent with the initial levels of trust as time/drive continued for each framing condition. This is predicted from the swift trust literature showing that the quick development of trust continues to hold because the automation's role has been made clear, so calibration follows the set path once the level of trust has been established (Greenberg et al., 2007; Kroeger et al., 2021).

2.1.2 Method

2.1.2.1 Participants. A total of 84 participants (61 female, 23 male; age *M* = 21.05, *SD* = 6.34) were recruited through SONA, an online research participation system, at Old Dominion University (ODU) during the 2022 Fall semester. All participants received credit towards a course research experience requirement. A demographics form was used to gather their information before the study (see Appendix A). This and the following studies have been approved by the Institutional Review Board at Old Dominion University according to the ethics guidelines.

2.1.2.2 Materials. The study was presented through a simulated driving environment created in STISIM driving simulation software (stisimdrive.com). The study was presented on a Dell P2717H 27-inch monitor with a 1920x1080 resolution and using a Logitech G29 wheel.

The study also employed the Trust in Automation Questionnaire (Jian et al., 2000). This survey has been used and evaluated in numerous studies and is the most well-known and utilized trust survey to date. The authors validated the survey through three experiments to develop an empirically-based scale. Though this study yielded terms relevant to trust and allowed for a 12-item scale, further validation of the psychometric properties was necessary. Spain and colleagues (2008) went on to further validate the questionnaire, finding from prior use of the study the internal consistency was acceptable (Fallon and colleagues (2005) had an internal consistency reliability of Cronback’s α = .93 and Safar and Turner (2005) additionally demonstrated its high reliability); however, they went on to test the validity for real-world use. Safar and Turner further validated, through confirmatory factor analysis based on their own experiment, that trust and distrust are related factors, which supports empirical findings from Benamati and colleagues (2006) and other real-world applications of trust (Balfe et al., 2018). Additionally, researchers have compared these ratings to trust-related behaviors like automation reliance and physiological measures (Bethel et al., 2007; Meyer, 2004; Montague et al., 2014). Based on these prior studies, this scale adequately assesses real-world trust in systems due to prior tests of reliability and validity, thus making it useful for the current study.

The drives used in this experiment were balanced among each other to be similar in terms of content regarding what participants experienced on the road. Typical experiences during driving such as merging vehicle, traffic lights, stop signs, construction events, etc. were included in the drives and balanced so that similar numbers of these events were experienced across drives. However, each drive was intended to be unique, so each drive contained different interactions at different times and unique environments. On average, the drives were six minutes.

Drive 1 was a two-lane road in a rural setting with a speed limit of 55 mph. It had various curves and hills, with one traffic light that it stopped at, several lane changes because of a slowed vehicle, and a brief section where the road was expanded to four lanes and then cut back down to two where the ego vehicle – vehicle that the driver is in – was forced to yield to a vehicle merging in.

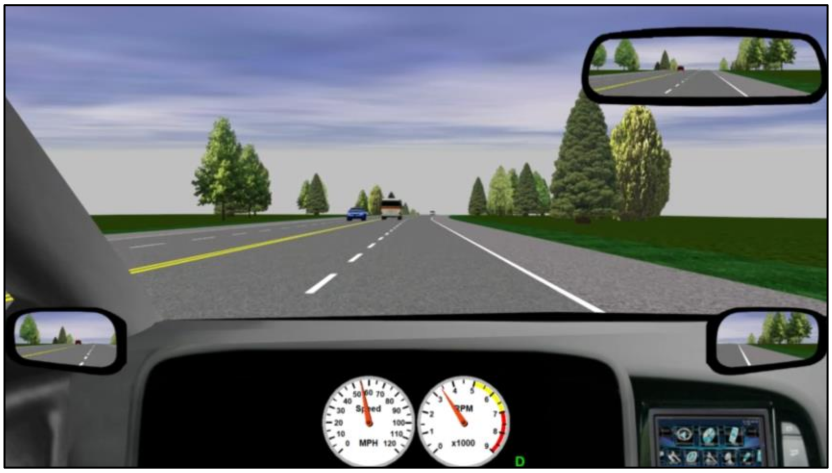
Drive 2 began on a six-lane highway at 55 mph which looked like it was exiting a city. Traffic merged to four lanes and the ego vehicle had to yield as it merged. Traffic merged down to two lanes and the ego vehicle had to zipper merge. There were also various curves and hills with a speed limit decrease due to a curve then back to 55 and had one traffic light.

Drive 3 started at 45 mph on a two-lane road in a rural area with several trees. The ego car had some cones on the right side of the road which it needed to merge slightly over into the opposing lane to avoid. There was one stop sign it had to stop at, then resumed driving and soon after the road expanded to a four-lane road with small buildings on the side and a few cars parked, with one car entering then exiting the roadway. There was one traffic light that it slowed for but then the light turned green.

2.1.2.3 Experimental design. The independent variables (IVs) included framing (positive/promotion, negative/dampening, control) and drive (Drives 1-3). Framing was the framing of the update information, which was manipulated between subjects. Participants were assigned to one of the Positive, Negative, or Control groups. In the instructions displayed on the driving simulator screen directly before Drive 1, participants were told to read the instructions – including the framing information – aloud to ensure that they fully read and understood them. The Control group presented with a neutral sentence, "For the next few drives, we have upgraded your vehicle with an automated driving feature". The Positive group was given the same base information as the control group with the additional positive frame, "We are excited about this new feature, and we are sure that you will enjoy it" and the negative frame was alternatively given, "We are testing something new, and you might need to keep watch to ensure everything goes well". The positive instructions were intended to show the potential gain from using the system (joy and help driving) and the negative instructions were intended to show the potential loss (unease and increased vigilance) After the instructions, participants went straight into the first drive so that their trust was not influenced by any practice, especially as this was framed as a new update. The factor of drive was manipulated within subjects, and participants experienced three separate and distinct six-minute drives. Figure 2 shows an image of one of the drives. All the drives contained similar experiences with various scenery, traffic, curves, etc. for the ADS to navigate, but were all uniquely different to show that they were different driving situations. The automation trust measure was the dependent variable (DV) and was administered after each of the three drives to assess their trust calibration over time.

Figure 2

A Screenshot of a Drive that the Participant Would See During the Experiment.



2.1.2.4 Procedure. Participants signed the consent form, were randomly assigned to one of the three framing conditions, and then read the instructions according to their group as they were seated at the driving simulator. They were told that the vehicle would start in the automated driving mode but that they needed to keep their hands on the wheel and foot near the pedals throughout the drive. Each drive was about 6 minutes long, with a total of 18 minutes of driving plus the transition period between each drive – no longer than one minute – for the survey and loading the next drive. During the drive, participants needed to monitor the performance of the ADS and be ready to take over if necessary, according to Level 2 automation requirements. However, they were not required to take over at any point as the performance for each drive was perfect. Then, upon completion of each drive, participants filled out the automation trust survey. Demographics information (including age and gender) was collected after completion of all three drives.

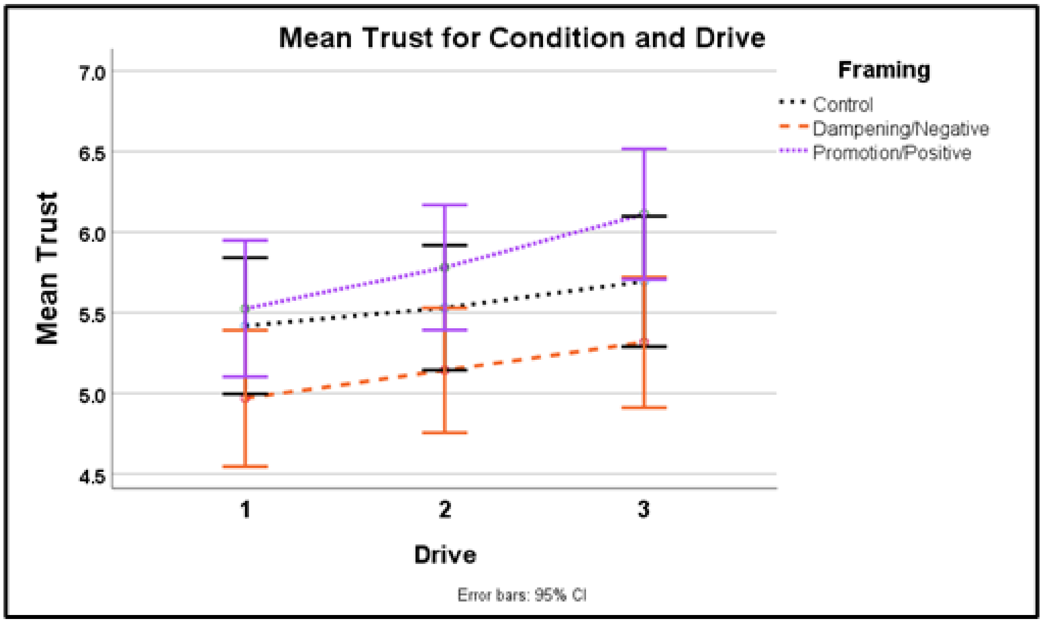
2.1.3 Results

A mixed Analysis of Variance (ANOVA) was conducted to determine the effect of the between-subject factor of framing (promotion/positive, dampening/negative, control) on trust over the within-subject factor drives (1-3). Analysis of trust showed a significant main effect of drive (*M*s = 5.30, 5.48, 5.71, for each drive 1-3, respectively), *F*(2, 162) = 11.29, $p < .001$ $\eta_p^2 = .12$. A linear trend analysis showed significant changes for between within-subject variables to determine increasing or decreasing trends. There was a significant linear trend of trust, *F*(1,81) = 16.67, $p < .001$, $\eta_p^2 = .17$, showing that trust significantly increased over time between drives. This supported H1 that trust would improve over time. See Figure 3 for graph.

The analysis of trust also showed a significant main effect of framing (*M*s = 5.55, 5.14, 5.81, for control, dampening, promotion, respectively), *F*(2, 81) = 3.27, $p = .043$, $\eta_p^2 = .08$. Pairwise comparisons using *t*-tests indicated that the mean score for the dampening condition (*M* = 5.14, *SD* = 1.23) was significantly lower than the promotion condition (*M* = 5.81, *SD* = 1.02), $p = .013$. However, the control (*M* = 5.55, *SD* = 0.92) did not significantly differ from the dampening or promotion conditions, *p*s > .05. Simple main effects analysis showed a significant difference between the dampening and promotion conditions for drives 2 and 3, *p*s = .023 and .007, but not for drive 1, $p = .063$. This provided partial support for hypotheses 2 and 3 that trust would be higher for promotion/positive than dampening/negative framing, and that these would continue to show differences as time went on due to the established framing and role. There was no significant interaction between drive and framing for trust, *F*(4, 162) = .625, $p = .645$ $\eta_p^2 = .02$.

Figure 3

A Graph of the Mean Trust Measure as a Function of Both Drive and Framing Condition.



*Note. T*he error bars are 95% Confidence Intervals.

2.1.4 Discussion

This study sought to test if trust calibration could be manipulated through something as simple as the way an update to an ADS was framed. The framing of a question can influence an individual's choice even when the resulting answer is the same. Trust calibration is a critical consideration for human interactions with ADSs because overtrust could cause the human to overestimate the capabilities of the automation and reduce their monitoring, resulting in the human not being in-the-loop when needed. In contrast, undertrust could cause the human to disengage the automation and not gain the benefits or try to overtake control at an improper time.

This experiment confirmed what prior studies of trust calibration in driving have shown regarding similar growth of trust over time through drives, demonstrating that experience with the system improves trust (Kraus et al., 2019; Mishler & Chen, 2023). Experiencing how the ADS performs through several drives without error improves trust and the current study showed that all framing conditions had similar increases in trust over time. However, the initial starting points and subsequent ending points showed that the framing of the automation update influenced trust. As expected, trust was lower for the dampening group than the promotion group. Note that for the initial drive the difference between the two groups was not significant but it was trending in that direction. However, even though trust improved over time, trust was significantly lower overall compared to the promotion condition for drives 2 and 3. This result supports previous and expected findings where some errors might be expected so the system provides transparency to the user through dampening framing to encourage proper trust calibration (de Visser et al., 2020). The trust dampening framing encourages the user to expect potential violations of trust so that if an error occurs, they can be ready. The transparency provided by the trust dampening also encourages an accurate mental model of trust during swift trust development.

The results of this study demonstrate that simple framing that either presents the information in positive or negative frame while introducing an ADS update can significantly influence an individual's trust. This helps to reinforce the previous literature that framing can change one's views even when the situation is the same, and that framing has the potential to influence an individual's trust calibration (de Visser et al., 2012; Dzindolet et al., 2002; Tversky & Kahneman, 1981). The present study showed that when trust was promoted, trust was created quickly compared to intentional dampening of trust. The negative frame was successful in dampening the trust calibration over time and is useful in situations where the automation might still be prone to errors and requires the individual to keep watch. Trust promotion can quickly build trust in the system, though if the system's capabilities do not match the trust, it would leave individuals with improper trust calibration. Similarly, if trust is dampened in a system that performs very well, individuals might not have enough opportunity to calibrate their trust through experience before disengaging from the system due to low trust. It is important when these new updates are rolled out to set proper levels of trust for the users so that they are calibrated in the loop as soon as possible without need for excessive experience.

Transparency helps clarify the ADS's role to the human driver and sets a realistic standard (Greenberg et al., 2007; Kroeger et al., 2021). Transparency itself using promotion or dampening may not be enough if the performance of the system does not match the information provided. As Greenberg and colleagues (2007) stated, it is not necessarily the role that matters, but how one performs in that role. Promoting trust is clarifying the role and trust is built very swiftly but it could fall greatly after a violation of trust. Thus, promotion may be improperly calibrating trust which could result in steep trust drops after a violation. Proper trust calibration is important, especially for human interaction with automation because humans are prone to automation bias, leading to unrealistically high standards for automation (Dzindolet et al., 2003). Because of the automation bias, people tend to react more negatively to automation failures as compared to humans who make the same mistake (Goodyear et al., 2017).

Properly clarifying roles and expectations for automation can be even more important in human-automation teams because of the biases that exist and a designer's ability to choose what information to show. By setting the expectations prior to interaction, trust can be swiftly developed in a trust promotion setting or dampened to realistic standards when the automation's performance might not live up to its expected standard. However, further research still needs to be conducted to examine how system errors can influence trust calibration while using promotion or dampening methods. Mishler and Chen (2023) showed significant drops in trust after an automation error, as did Kraus and colleagues (2020). There was a natural repair of trust after the violation; however, neither study included transparency information about the systems' capabilities or role. De Visser and colleagues (2020) estimate that promotion and dampening can help influence proper calibration to enhance trust resilience in situations of error, but more work still needs to be done in this area.

One limitation of this experiment is that we did not include an initial or dispositional trust measure before the study. A premeasure of trust would ensure that all participants were starting on the same level of trust. However, with 28 participants per group, any potential extremes in beginning levels of automation trust between groups should regress to the mean so the groups should be even in terms of initial trust. Additionally for this initial investigation, we did not want to prime individuals about trust before the framing manipulation by asking them about trust in automation. Priming them with trust might make them dwell on the idea of automation trust thus rending the framing manipulation less effective, decreasing external validity.

Another limitation is that this experiment featured no automation errors and the vehicle performed perfectly for each drive. This was intended as a baseline to understand how both promotion and dampening framing might affect trust in an ADS update scenario. The next study includes an error during the driving sequence to see if framing makes trust more resilient and if it effects the repair of trust after an error.

2.2 Experiment 2

Experiment 2 built upon the findings from Experiment 1 and implemented an automation error into one of the drives for two purposes: (1) to further validate the results on how framing impacts initial trust calibration and (2) to examine how framing affects trust calibration following an automation error. We also wanted to ensure that all participants had the same trust and opinions on automated driving coming into the experiment, which was a potential limitation from Experiment 1. However, the goal of Experiment 1 was to ensure that the framing effect was viable and any questions about trust in automation beforehand might prime them about trust making the initial framing less potent. With Experiment 2 we controled for the initial trust and attitudes toward automation to understand how framing influences trust after errors. Framing should help to form the trust calibration to make sure the operator responds to automation errors in a safe manner.

2.2.1 Hypotheses

Following Experiment 1, we expected some initial differences in trust between framing conditions after the second drive. However, as stated above, the presence of the initial trust and opinions towards automation survey at the beginning of the study might influence the initial effect of the framing by causing participants to consider the idea of trust before reading the prompt and before the first rating of trust. Therefore, we expected trust to show some differences in the initial trust for Drives 1 and 2 with promotion showing higher trust scores than dampening (H1), with the caveat that Experiment 1 did not show a significant difference for Drive 1.

We expected that after the error, trust will decrease for all conditions regardless of the framing (H2). We do not expect differences between the decreases in trust across the conditions because errors perceived as a failure or major violation will decrease trust across the board, regardless of framing, as has been found repeatedly in trust research (Hoff & Bashir, 2015; Lee, & See, 2004; Mishler & Chen, 2023; Muir & Moray, 1996). Previous research has found errors cause similar decreases in trust even when varying task difficulty, error type, and reliability (Kraus et al., 2019; Mishler, 2019; Mishler & Chen, 2023).

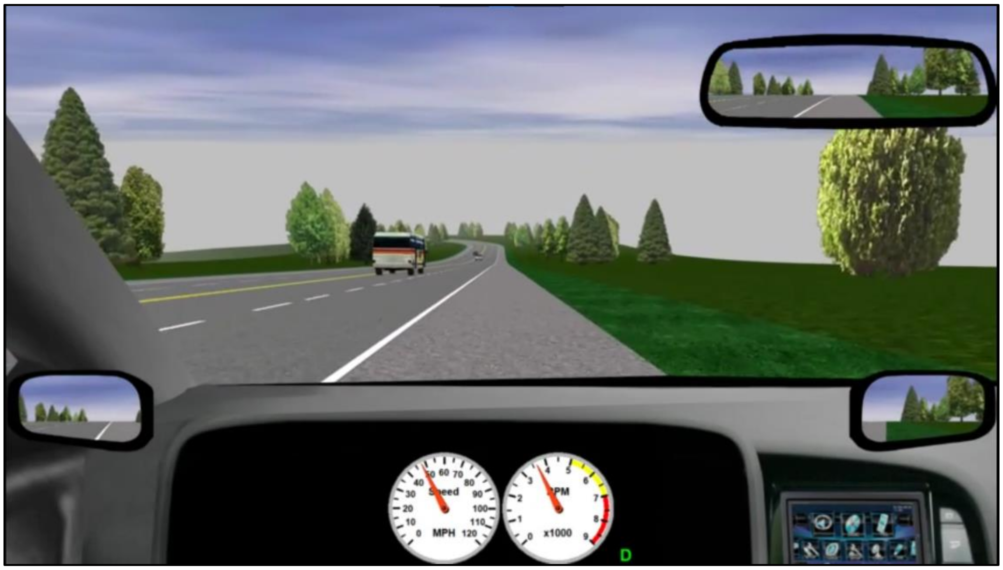
We expected the trust recovery to be different for the framing conditions for those drives after the error (H3). Trust should be higher for the trust promotion condition than for the trust dampening condition because stronger swift trust has already been established earlier on. (Haring et al., 2021; Kroeger et al., 2021; Meyerson et al., 1996). The dampening group did not have the bolstered trust and should have lower trust afterward, but this ought to be an expected calibration of trust after an error (de Visser et al., 2018, 2020; Kraus et al., 2019; McGuirl & Sarter, 2006).

2.2.2 Method

The method for Experiment 2 was similar to Experiment 1 except that the number of drives was increased from 3 to 5 to include the error on Drive 3 and the participants were given the adapted Automated Driving Opinion Survey (ADOS; Kyriakidis et al., 2015; see Appendix B) at the beginning of the study. The third drive included an automation error (see Figure 4) in which the ADS does not stay in the lines of the road and swerves off the road for several seconds, driving partially on the shoulder. After all five drives were completed, the researcher conducted a structured interview of around ten questions (see Appendix C) asking about participants’ experience with the system, if they noticed the error, and what they thought about the error and how that reflected on the system.

Figure 4

The Driving Error with the Vehicle in the Shoulder During Drive 3 in Experiment 2



2.2.2.1 Participants. We ran 72 participants from the ODU SONA pool based on the G*power analysis using an estimate of $\eta_p^2 = .08$ from the main effect of framing in Experiment 1, resulting in effect size $f = .29$. This calculation showed that for a power of .80 the total sample size should be around 72 participants (see Appendix D). Therefore, we collected 72 participants (24 in each group). The mean age was 20.17 (*SD* = 4.76) years and there were 17 males and 55 females. The participants were required to have a valid driver's license and reported that their amount of driving experience was 3.63 (*SD* = 1.49). The participants' ethnicities were Caucasian (25), African American (30), Asian (3), American Indian/Alaska Native (2), more than one race (8) and other (3). All participants had normal or corrected-to-normal hearing and vision. The students recruited from SONA were given class credits.

2.2.2.2 Materials. The Automated Driving Opinion Survey (ADOS) was based on Kyriakidis and colleagues' (2015) survey but was be modified to use only questions about driver behavior and attitudes toward automated vehicles. It was implemented to gain insight into participants' initial thoughts on automated driving before beginning the study. The ADOS helped assess whether participants in each group are similar in their thoughts and opinions toward automation. This survey was used to gather manual driving experience, automated driving experience, and opinions on automated vehicles from each of the participants. Kyriakidis and colleagues (2015) used the survey to find out about public opinions on automated driving across many cultures and therefore contains many unrelated questions, such as income and disability. Therefore, the participants were only be asked the questions related to manual and automated driving behaviors and opinions. This scale was given to participants at the beginning of the study and showed overall information about participant's beliefs about autonomous vehicles before they perform any of the automated driving in the experiment. The results of the ADOS were used as a covariate to control for potential differences on these items.

The drives in the study were the same as those from Experiment 1, but a new drive 3 was inserted so that the error could occur, making the former drive 3 now drive 4. Additionally drive 5 was added so that there were now two initial drives with no error, an error on the third drive, and two more drives after with no error.

Drive 3 was a four-lane highway with a shoulder and a speed limit of 55 mph. The ego vehicle had to navigate curves and hills and two vehicles on the shoulder at separate times, with the second one merging onto the road. Later towards the end of the drive, the error occurred where the ego vehicle crossed the solid white line on the right side of the road and rode in the shoulder for 10 seconds, attempted to return to the lane only making it halfway before returning to the shoulder, then fully recovering back onto the road. This was done to make it look like the ADS had lost the main road lane and was navigating its way back on.

Drive 5 started at 35 miles per hour as it was driving out of a neighborhood two-lane road with several houses and parked cars. Once it left the neighborhood it sped up to 45 mph on a rural road with many trees. It came to a small city with one traffic light that it must stop at before continuing back to the forested road. The road expanded to a four-lane road then later the ego vehicle had to slow down to 35 mph for a construction event with one non-moving, seated worker and it was not required to change lanes in the construction zone. After the construction it speed back up to 55 mph.

Participants were also interviewed after the study with a brief structured interview (Appendix E) to understand what their subjective thoughts were about the automation during the study and how they thought about the error. This allowed them to elaborate more on their trust in and opinion about the automated vehicle giving more qualitative data to help explain their behavior and inform the design of Experiment 3.

2.2.2.3 Procedure. Participants first signed the consent form and be randomly assigned to one of the three framing conditions (promotion, dampening, control), and then read the instructions according to their condition after they are seated at the driving simulator. Similar to Experiment 1, participants were instructed to read the instructions aloud before Drive 1. They were told that the vehicle would start in the automated driving mode, in which the vehicle can control lateral and longitudinal movement, but they needed to keep their hands on the wheel and foot near the pedals throughout the drives. During the drives, participants needed to monitor the performance of the ADS and be prepared to intervene if necessary, according to Level 2 automation requirements. They performed five drives in total but were not told how many there would be ahead of time. Then, upon completion of each drive, participants filled out the automation trust survey. Near the end of Drive 3, an error was introduced, and the vehicle slowly drifted over the right road line into the shoulder. The vehicle fully crossed the line, drove halfway back as if trying to find the road, drove further right again before finally correcting and driving back to the center of the lane. This entire error process took approximately 20 seconds. They performed five drives in total but were not told ahead of time in terms of how many there would be or any other information about them, including the upcoming error. Demographics information (including age, gender, driving experience) was collected after completion of all five drives. After the driving simulation section was completed, the instructor conducted the structured interview, digitally recording their responses via a recording tablet. After the conclusion of the interview, the experiment was completed and they were thanked, given credit, and ushered out of the lab.

2.2.3 Results

2.2.3.1 Quantitative Results. This section contains the quantitative analyses of the ADOS and trust data collected prior and during the experiment, separate from the qualitative data from the post-experiment interview.

To confirm that there were no differences between groups for initial trust, ADOS was analyzed using a one-way ANOVA with condition (control, dampening, promotion) as the between-subjects independent variable. The analysis revealed that there was not a significant effect of ADOS between condition (*M*s = 4.56, 4.69, and 4.68 for control, dampening, and promotion, respectively), $F < 1$. Follow up equivalence tests for the ADOS data using the two one-sided t-tests (TOST) method to verify there were no differences for starting automated driving opinions between groups, further clarifying that all groups started equal. The TOST test used 90% confidence intervals (CIs) around the mean and rejects any effects of Cohen's $d = 0.4$ or larger to ensure that if the CIs are within the effect size interval, we can conclude equivalence between the two samples (Lakens, 2017). We used two one-sided tests – right and left – to obtain *p* values tested against the defined interval, with the right one-sided *t*-test on the lower bound and a left one-sided *t*-test on the upper bound. The greatest *p*-value reflects that of the equivalence test, and a significant effect means a claim of equivalence is supported. All compared groups' 90% CIs for equivalence was within the equivalence interval of [-.5, .5], *p*s < .05, meaning a claim of equivalence is supported. Table 1 shows the results of the *t* tests with the greater of the two *p* values listed for each group comparison.

Table 1

Results of the TOST Equivalence Test for the ADOS in Experiment 2

	Control	Promotion	Dampening
Control	-	*t*(46) = 2.19, *p* = .017	*t*(46) = 1.79, *p* = .040
Promotion	*t*(46) = 2.19, *p* = .017	-	*t*(46) = 2.25, *p* = .015
Dampening	*t*(46) = 1.79, *p* = .040	*t*(46) = 2.25, *p* = .015	-

Note: Significant effects indicate that the groups are equivalent

To test H1-H3, I analyzed the trust data similar to Experiment 1, but used a 3x5 mixed Analysis of Covariance (ANCOVA) with Framing (Control, Promotion, Dampening) as the between-subject variable, Drive (1-5) as the within-subject variable, and the automated driving opinion survey (ADOS) results as the covariate to account for differences in initial trust and opinions on automated driving systems. Even though there were no differences between groups for initial trust, individuals within the groups might have different starting levels of trust and adding the ADOS as a covariate helps to control for these differences.

Of note, I used Type 2 Sum of Squares for these analyses to avoid throwing out a portion of model power. Type 3 Sum of Squares is more often used; however, the benefit of using Type 2 in these analyses is that it considers the interactions with higher importance, but also recovers the power for the main effect if interactions are not present (Langsrud, 2003). Due to the inclusion of multiple drives and the covariate, this was the most viable consideration for the analyses.

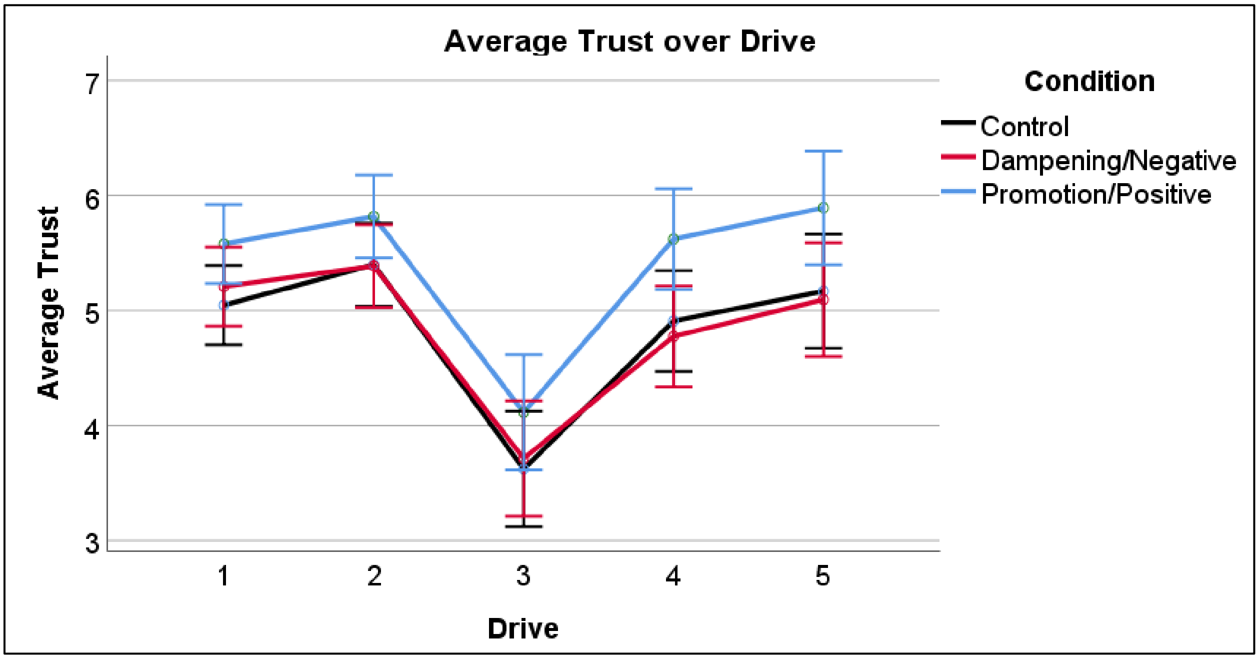
The main effect of drive was significant, $F(4, 272) = 75.01$, $p < .001$, $\eta_p^2 = .52$, indicating that trust was not the same across drives (*M*s = 5.28, 5.53, 3.82, 5.10, 5.38 for drives 1-5, respectively). See Figure 5 for graph. This main effect of drive helps support H1, H2, and H3, demonstrating that there are changes in trust at different points in time. To investigate the differences across the drives, a trend analysis was used to analyze the overall pattern of trust using within-groups contrasts. The analysis revealed a significant quadratic trend, $F(1, 68) = 11.09$, $p < .001$, $\eta_p^2 = .14$, which demonstrated a trendline with one significant curve, indicating a decrease in trust after the first two drives because of the error in drive 3, then an increase in trust for the final two drives. The trend demonstrates the significant decrease in trust for the third drive compared to the initial and subsequent errorless drives. The trend analysis provided support for H2, demonstrating that all conditions decreased during the error in drive 3. There was a significant effect of the covariate ADOS, indicating that an individual’s prior opinions on automated vehicles influenced their trust during the study, $F(1, 68) = 6.65$, $p = .012$, $\eta_p^2 = .09$. Through the use of ADOS as a covariate, the analysis was able to help control for initial levels of trust and opinions of automated driving, eliminating preconceived ideas about automation as a factor affecting the data.

There was a significant main effect of framing, *F*(2, 68) = 3.45, *p* = .037, η_p^2 = .09. Planned contrasts showed that participants in the promotion framing condition (*M* = 5.40) were significantly more trusting than those in the control condition (*M* = 4.83), *t*(68) = 2.29, *p* = .025, and those in the dampening condition (*M* = 4.84), *t*(68) = 2.26, *p* = .027. The planned contrasts provided further support for differences of the promotion framing condition, demonstrating that the promotion condition was significantly higher in trust than the other conditions.

There was a significant interaction between drive and framing for trust, *F*(8,272) = 2.11, *p* = .035, η_p^2 = .05. Simple main effects analysis showed a significant difference between the control and promotion conditions for drives 1, 4, and 5, *p*s = .033, .025, and .043, but not for drives 2 or 3, *p*s = .106 and .171, respectively. The promotion condition was also significantly different from dampening for drives 4 and 5, *p*s = .008 and .026, but not drives 1-3, *p*s = .133, .096, and .263, respectively. There were no significant differences between control and dampening conditions, *p*s > .05. This provided partial support for H1 and H3 that trust would be higher for promotion/positive than dampening/negative framing, and that these would continue to show differences after the error due to the established framing and role. There was no significant interaction of drive and ADOS, *F*(4, 272) = 1.64, *p* = .164, η_p^2 = .02.

Figure 5

Average Trust Over the Five Drives for Each of the Three Conditions in Experiment 2



Note. The ADOS covariate appearing in the model was evaluated at 4.64 and the error bars are 95% Confidence Intervals.

2.2.3.2 Post-task Interview Results. This section contains the qualitative analyses from the post-experiment interview. Voice recordings of the interviews were fed into an AI transcription service (dovetail.com) to generate text transcriptions, and then undergraduate research assistants, and I went over the transcriptions to ensure the accuracy. The data was then fed into MAXQDA (maxqda.com), a qualitative data coding tool. The interview data was coded to mark each question's response. For any keywords including mentions of trust, safety touching wheels/pedals, vehicle performance, the relevant statement was marked as well as other interesting statements up to the coder's discretion. The coded statements and keyword phrases were separated, analyzed for their respective important information, and is reported below.

The interview data showed a variety of results. For the first question (Q1) that participants were asked after being reminded of the initial framing instructions, "When you read this instruction at the beginning, what did you think about the vehicle?" we expected more comments about the actual framing text but many of the participants only commented on their expectations of what the vehicle would be doing. We did not want to influence them by asking directly if the instructions gave them a positive or negative impression of the system because that was too leading. However, the resulting responses to this question were too varied and broad to understand their direct thoughts about introspective feelings toward the automated system after the instructions. In general, participants still felt uncertain about how the vehicle would perform or behave before starting the drive.

Participants were asked if they noticed anything strange about the automated driving system during their experience (Q5) and a vast majority (67 out of 72 or 93.06%) of them mentioned the error of driving off the road either during this question, or at some point prior to the question being asked. The remaining participants that did not mention the error during this question or prior to the question were shown a short clip of the error and asked if they had seen this occur and all did see it and through later questioning did consider it an error of the automation. They had either forgotten to mention it or described it vaguely in previous statements that were not fully clarified. This question helps verify that participants were noticing and responding to the error and that it was noteworthy enough to be acknowledged.

The seventh question (Q7) asked participants, "What did you think might be the reason for the error?" and the main responses to this question were either a "glitch in the system", a "sensor or lane issue", or an "AI or system training issue". The respondents who mentioned that it was just a glitch in the system demonstrated an overall lack of knowledge about how these automated system work, which was to be expected. They did not go into any detail about what specifically might be wrong, but essentially said that technology is just broken sometimes, and that kind of stuff happens. Some individuals mentioned the sensor or lane detection issue which shows a more specific understanding of what piece of the technology could have been causing the error. These participants showed a more critical understanding of the issue, applying a better knowledge of how ADSs behave and rely on sensor and other input to process the road and drive. Lastly, a few participants showed an understanding of how AI is trained to understand and react to driving environments through training data. They mentioned that since the system might be new it did not have enough training or was not trained on enough roads of this type so it might need further fine tuning. This question shows the individual differences in knowledge bases and mental models of an ADS, which could influence their trust development in the system differently, prior to an experience, during an error, and through later interaction. Unfortunately, there was not enough data collected to show a distinct difference between groups but all the individuals with different understandings of the error were equally spread throughout the priming groups.

Participants were directly asked (Q8) if the error "influenced their trust in the vehicle's capabilities" and they all reported that their trust decreased, albeit to a varied extent. Some reported that it only affected their trust a little bit, whereas others stated that it brought their trust way down. Much of this data is already in the questionnaire, but of note from this question is some of their explanations of why their trust decreased or what happened after the error. Several of the participants mentioned thoughts along the lines of other potential errors that might occur given the error they just saw happen during the drive, encapsulated in this response from a participant, “It would affect me really, because I don't know what else like it can do like error can do if there's like an accident or something, right? Like if there's like a kid just suddenly ran right in the street and I don't know if the car is gonna be fully, you know, fully stopped or, or keep going.” However, several other participants noted that while they were concerned with the error in the moment, the automation was able to correct itself and then even later on it drove well enough to make them trust it again, which can be seen in this response, “It shook it a little bit, not gonna lie. But then after that we were presented with a lot more, you know, stuff in the road and cars coming out in front of us that the system was able to correct. So, it made… it restored my confidence in the safety.” This again highlights the different emphasis individuals place on various aspects of the system and what their attitudes and opinions can be like depending on their priority and mental model.

Participants had a diverse reaction when asked (Q4), “If you had the chance in real life, would you use the automated vehicle that would perform the same as the one you just saw in the study?” There were 40.28% (29) of participants who said yes, 43.06% (31) of participants who said no, and 16.67% (12) of participants who were uncertain or needed more experience with the system first. Several mentioned a mixed opinion stating that they might be interested in checking it out, but they would not want to give up full control or that using the system would be boring or takes the fun out of driving and they would rather drive themselves. However, others mentioned that it would be nice to use sometimes and for certain situations but not as a use in everyday situations.

2.2.4 Discussion

Experiment 2 investigated the framing effect's ability to influence swift trust before interaction with an automated driving system experiencing a new update to fully automated driving while also experiencing an error. The outcome builds upon the findings of Experiment 1 which found that manipulations of how an update was described to an individual can influence their initial trust and later trust development. We found similar results for the initial levels of swift trust after their first interaction, then further supported the hypotheses that a positive framing of the automated system would lead to higher calibrations of trust over the subsequent drives. The critical new finding of this study was that after receiving a positive trust framing, participants continued to have higher trust even after an error.

Similar to Experiment 1, framing the update instructions in a positive way contributed to a higher calibration of swift trust and then subsequent trust levels later on. However, Experiment 2 demonstrated increased trust even after an error for the trust promotion group. Even though the automation performed exactly the same for each group, those receiving a positive framing before interaction still had higher trust recovery after an error, showing a potential for greater trust resilience and better trust repair. Trust after the error in the third drive was not significantly different between framing conditions but did show a trend that the trust promotion condition was higher. Trust recovery after the error in drives 4 and 5 further demonstrated that trust promotion at the beginning stages before user interactions will continually influence trust over time during the interaction. The interactions with a system is often cited as a highly influential factor for building and calibrating trust (Hoff & Bashir, 2015; Kraus et al., 2019; Metcalfe et al., 2017) and is typically how users discover the reliability and capabilities of the system. However, even though the users were interacting with the exact same system, their trust was indeed manipulated prior to interaction which influenced later trust.

The design of the framing for both the promotion and dampening conditions was to influence their initial swift trust by importing specific capability information and telling them about the role (Kroeger et al., 2021; Meyerson et al., 1996). The promotion condition focused on reassuring them that the automation was good (importing) and would take care of everything (role) whereas the dampening let them know it might need oversight (importing) and they need to keep watch (role). Kroger and colleagues (2021) found positive associations between role clarity and swift trust; however, Experiment 2 only found higher swift trust for promotion. This would be expected because the description of role for dampening told them they would need to oversee the automation and was intentionally phrased to not increase swift trust. This finding adds to swift trust literature showing that the type of role information given can influence swift trust and that simply clarifying role might not increase trust. Similarly with importing trust, Kroeger and colleagues (2021) found a positive association between positive reputational information and swift trust. Trust promotion highlighted the positive reputational aspects of the system, imbuing confidence in the system’s abilities, whereas the dampening attempted to make the individual more uncertain about the capabilities. The promotion was able to increase swift trust and maintain that over time, whereas the dampening was similar to the control, showing that swift trust might be harder to dampen and might not exist on an opposite spectrum (trust increases might not function the same as trust decreases). This might also be a limitation of the trust measurement scale that does not fully capture negative aspects of trust.

Providing participants with too much information can limit their understanding and overwhelm them. A balance of detail given to the individual, and the length of the content is required. Ideally, giving individuals all the information about the system to make a cohesive mental model of the system would allow for the best trust calibration. However, even that is not feasible because individuals could not maintain and process all that information and would develop different mental models. Therefore, it is best to make the information short and pertinent to the task (Brust-Renck et al., 2013; Lacson et al., 2005).

A few outcomes from the interview data were that the error was sufficient to evoke a reaction from the participants because the majority noticed and mentioned it in the interview or did recall it after being reminded. I decided to keep the error the same for Experiment 3 based on these findings. Another outcome that showed some potential for individual differences based on users' understanding of the technology they are interacting with. The varied responses to the question asking about their understanding of why the error might have occurred demonstrate a differential understanding and mental model of how and ADS functions and what systems exist within automated driving. Similarly, participants had a varied reaction to the error, thinking about the potential for other more significant errors later on, or understanding that these small problems can occur, but the later performance reassured them that the system was okay. Though the interview data is broad and cannot fully explain participant behavior, it shows how various differences in experience to the same situations can result. Through further intervention I to evoke these different responses through instructions and system messages.

Given the findings from this study, I moved on to the final experiment of the series which investigated the same scenario but with the addition of active trust repair after the third drive.

2.3 Experiment 3

The goal of Experiment 3 was to investigate the effects of different active trust calibration strategies after an error to further influence trust calibration after initial priming. The initial impacts of framing on swift trust can help bolster the initial building of trust as shown in the first two experiments, but swift trust might have implications for trust behavior in the case of an error even after extended interaction. Based on de Visser and colleagues (2020), there are multiple ways to either repair or dampen trust after an error. The intent of repair would be to continue promoting trust to the individual, therefore the participants in the trust promotion condition were presented with a trust repair strategy after the error. One of the methods most for **trust repair** would be an apology where fault is admitted, and some form of regret is conveyed (Kohn et al., 2018; Nayyar & Wagner, 2018). Kohn and colleagues used a variety of different strategies that have been successful in human-human studies and showed that two of the strategies involving apology (timed apology and apology with entity attribution) showed the most success. The apology can be conveyed with the addition of a promise to improve in the future. Modern systems are constantly adapting, and telemetry sent to the manufacturer or even AI processes could help ensure similar mistakes are actually decreased in the future. However, some cases might not be easily fixable, and a dampening strategy might be more useful. The goal of a dampening strategy is to calibrate trust to be lower, preventing overtrust. Therefore, participants in the dampening condition were presented with a trust dampening strategy after the error. De Visser and colleagues (2020) again put forward a method for **trust dampening** as conveying the system limitations. The operational domain is highly important in the field of automated driving and certain systems cannot function if certain conditions are not met (SAE, 2021). For example, faded lines on the road surface might cause an LKS to veer away from the middle of the lane. If the vehicle notifies the individual of this limitation, it could dampen the user’s trust in the system, but improve their trust calibration for the system leading the human to keep a better lookout for such situations in the future.

For Experiment 3, I implemented the apology-with-promise (trust repair) and conveying-system-limitations (trust dampening) strategies after the error in Drive 3 to the trust promotion and trust dampening groups, respectively. The intent of combining multiple repair and dampening strategies (e.g., apology and promise, and error explanation with limitations) was to get the highest effect of the intervention. This is based on prior advice on trust repair research that states no strategy should be implemented by itself and should be combined with others for the most benefit (de Visser et al., 2018). The goal was to test the implementation of an active trust calibration strategy after an error in addition to the framing manipulation from the beginning of the experiment. The results were also compared to Experiment 2 using a quasi-experiment design to examine the effects of the active repair/dampening strategy after an error.

2.3.1 Hypotheses

I expected a main effect of framing-with-active-calibration (H1), such that trust repair (with both promotion framing and active trust repair after the error) would have higher overall trust compared to the trust dampening (with both dampening framing and active trust dampening after the error). By their nature, repair strategies are intended to prevent decreases in trust and dampening is intended to keep trust lower (de Visser et al., 2018, 2020; Kohn et al., 2018).

I expected results similar to Experiment 2's H1, trust should be higher for promotion group showing higher trust scores for Drives 1 and 2 than dampening (H2), with the caveat that Experiment 1 did not show a significant difference for Drive 1.

I expected that following the error and trust calibration strategies in Drive 3, trust would be highest for the promotion, showing little trust decrease, moderate trust decreases for dampening as they were already calibrated for anticipating errors, both of which should show smaller decreases than the control as the system does not acknowledge the error in the latter

(H3). The initial framing at the beginning should work as a preventative measure for the trust dampening condition because they already expected a potential trust violation as was explained through the potential limitations included (de Visser et al., 2020). The control condition had no such indication and only saw the error, counting as a miss that the human noticed, which is known to decrease trust (Hoff & Bashir, 2015).

I expected that for Drives 4 and 5, trust would continue to increase for the promotion condition, would slightly improve for dampening, and would rebound for control (H4). Trust dampening should more appropriately calibrate trust – less trust increase after an error – because role and expectation are already established and an error should not be a surprise (de Visser et al., 2020; Haring et al., 2021; Kroeger et al., 2021). For the control, trust violations decrease trust but show improvement after subsequent perfect trials (Hoff & Bashir, 2015; Kraus et al., 2019; Mishler, 2019; Mishler & Chen, 2023).

I expected that for the trust repair condition within Experiment 3, trust would be higher directly after and following the error compared to the promotion framing condition in Experiment 2 and that for the dampening condition in Experiment 3, trust would be higher directly after the error compared to the dampening framing condition in Experiment 2 (H5). Trust repair should prevent declines in trust and continue promoting trust even after an error (de Visser et al., 2020). Trust dampening after an error should still show an acknowledgement of the error compared to completely ignoring the error, which would provide some transparency. This should decrease trust but not as much as ignoring the error because the system primed them to beware that an error may occur and then explaining the system limitations as initially described (de Visser et al., 2020).

2.3.2 Method

The method for Experiment 3 was similar to Experiment 2 with the exception of three changes. The first change was that the initial framing for the negative/dampening condition was adjusted to be clearer and avoid using specific language telling participants to watch the road. This adjustment was done to ensure the negative/dampening condition was more consistent with the wording in the positive/promotion condition. The statement, “We are testing something new, and you might need to keep watch to ensure everything goes well” was changed to, “We are testing something new, and we are not yet sure that it will perform perfectly”. Additionally, the control condition was updated with, “We have completed the installation of the update, and you may proceed with your drive now.” as an additional line to ensure the length of each passage was equivalent. This addition did not impart any new information about the vehicle but gave a neutral sentence telling them they could now use the car. All initial framing instructions were balanced for word number, with a total of 26 words each.

The second change involved a few updates to the structured interview questions (Appendix F). I added a few more questions based on the addition of the promotion and dampening information provided to participants after the third drive. Additionally, a couple of questions were added to elicit more elaboration about the participants’ experience during the error. Lastly, the question about if the participant would use the vehicle in real life was added to the end.

The final and most important change was the addition of the active calibration using trust repair or dampening strategies implemented following the error of Drive 3. To stay similar to the initial framing, participants in the trust promotion condition saw an active trust repair message after the error and participants in the trust dampening condition saw an active trust dampening message after the error. At the end of Drive 3, participants in the trust promotion condition received the trust repair information as an apology-with-improvement. The apology-with-improvement was implemented as a visual dialogue box that they read aloud, similar to the initial display of trust promotion at the beginning. The trust repair message said, "I am sorry for leaving the center of the lane and driving on the road edge. I have processed the diagnostic information and will not make this mistake in the future". Participants in the trust dampening condition received trust dampening information after the error in a similar way, in the form of conveying system limitations and were told, "I left the center of the lane and drove on the road edge. The previous section of the road had irregular road lines, which is outside of my current system limitations." Participants were able to see the road line during the drive and there was no noticeable fading, though I did not expect participants to notice at the time because it would have happened at the beginning when the ADS first went off the road. The control condition was included to keep a message in the drive after the error to ensure it was balanced with the other two conditions. It was difficult to balance this condition so that it did not tip off the participant about the error or seem like an acknowledgement of the error. Based on the responses from Experiment 2, participants noticed the error by themselves – as expected – so the extra message should not inform them of anything additional and as long as the message does not discuss the error, it should not be directly tied to the error. So, for the message, it did not contain any information relevant to the drive or automated system, but said, "You have a new notification from American Red Cross which is reminding you to donate blood. There are several blood drives in your area if you want to schedule an appointment."

2.3.2.1 Participants. I recruited 72 participants (24 in each group) from the ODU SONA pool or ODU community based on the same power analysis from Experiment 2. The mean age was 22.56 (*SD* = 7.14) years and there were 26 males and 46 females. The participants were required to have a valid driver's license and reported that their amount of driving experience was 4.28 (*SD* = 1.90) years. The participants' ethnicities were Caucasian (31), African American (19), Asian (4), American Indian/Alaska Native (1), more than one race (13) and other (4). All participants had normal or corrected-to-normal hearing and vision. The participants from the ODU community were recruited via flyers posted on campus or a flyer link in the psychology daily announcements email and emails to psychology course teachers to forward to their students with the intent of capturing psychology student populations and result in either students, professors, or others employed by ODU. The students recruited from SONA were given class credits and the participants from the community were given $10 in Amazon gift cards for their time.

2.3.2.2 Materials. Similar to Experiment 2, The ADOS based on Kyriakidis and colleagues (2015) was implemented to gain insight into participants' initial thoughts on automated driving before beginning the study. This helped evaluate whether participants in each group were similar in their thoughts and opinions toward automation.

Participants were also interviewed after the study with a brief structured interview to understand what their subjective thoughts were about the automation during the study and how they thought about the error and trust calibration strategy. As stated in the method, this was similar to the structured interview from Experiment 2, but with a few more added questions (Appendix F).

2.3.2.3 Procedure. Participants signed the consent form and were randomly assigned to one of the three framing conditions (promotion, dampening, control), and then read the instructions aloud according to their group after they were seated at the driving simulator. In the instructions were displayed similar to the prior two experiments, requiring the participants to read the instructions aloud. They were told that the vehicle would start in the automated driving mode but that they need to keep their hands on the wheel and foot near the pedals throughout the drives. During the drives, participants needed to monitor the performance of the ADS and be prepared to intervene if necessary, according to Level 2 automation requirements. They performed five drives in total but were not told how many there would be or any information about them, including the upcoming error. Then, upon completion of each drive, participants filled out the automation trust survey. Near the end of Drive 3, the vehicle slowly drifted over the right road line into the shoulder. The vehicle fully crossed the line, drove halfway back so as if trying to find the road, drove further right again before finally correcting and driving back to the center of the lane. This entire error process took approximately 20 seconds. After the automation had recovered from the error, participants saw a dialogue box on the screen showing them the active trust calibration message, according to their condition. They were instructed to read this aloud to ensure they had fully read and understood the message. Demographics information (including age and gender) was collected after the completion of all five drives. After the driving simulation section was completed, the instructor conducted the structured interview, digitally recording their responses via a recording tablet. After the conclusion of the interview, the experiment was complete and they were be thanked, given credit, and ushered out of the lab.

2.3.3 Results

2.3.3.1 Quantitative Results. This section contains the quantitative analyses of the ADOS and trust data collected prior and during the experiment, separate from the qualitative data from the post-experiment interview.

To confirm that there were no differences between groups for initial trust, ADOS was analyzed using a one-way ANOVA with condition (control, dampening, promotion) as the independent variable. The analysis revealed that there was not a significant effect of ADOS between condition, (*M*s = 5.01, 4.77, and 4.73 for control, dampening, and promotion, respectively) *F* < 1. Follow up equivalence tests for the ADOS data using the two one-sided *t*-tests (TOST) method to verify there were no differences for starting automated driving opinions between groups, further clarifying that all groups started equal. Only the promotion and dampening group’s 90% confidence interval for equivalence was within the equivalence interval of [-.5, .5], *p*s < .05, meaning a claim of equivalence is supported. However, for the control and dampening and the control and promotion groups the results were not significant. Therefore, the conclusion for equivalence is inconclusive. The overall ANOVA (*F* < 1) and the individual independent samples *t*-tests between each set of groups, *t*(46) = 0.96, *p* = .343, *t*(46) = 1.37, *p* = .176, *t*(46) = 0.17, *p* = .864 for the groups control and dampening, control and promotion, and promotion and dampening, respectively). Therefore, the data cannot fully support a claim of equivalence, but does provide some evidence that they are not different. Based on these results, the ADOS was used as a covariate in follow-up analyses to ensure any differences are controlled. Table 2 shows the results of the *t*-tests with the greater of the two p values listed for each group comparison.

Table 2

Results of the TOST Equivalence Test for the ADOS for Experiment 2

	Control	Promotion	Dampening
Control	-	*t*(46) = 1.07, *p* = .145	*t*(46) = 1.06, *p* = .148
Promotion	*t*(46) = 1.07, *p* = .145	-	*t*(46) = 1.83, *p* = .037*
Dampening	*t*(46) = 1.06, *p* = .148	*t*(46) = 1.83, *p* = .037*	-

Note: Significant effects indicate that the groups are equivalent

To test H1-H4, I analyzed the trust data similar to Experiment 2, using a 3x5 mixed ANCOVA with Framing (Control, Promotion, Dampening) as the between-subject variable, Drive (1-5) as the within-subject variable, and the automated driving opinion survey (ADOS) results as the covariate to account for differences in initial trust and opinions on automated driving systems.

The main effect of drive was significant, $F(4, 272) = 49.92$, $p < .001$, $\eta_p^2 = .42$, indicating that trust was not the same across drives (*M*s = 5.21, 5.54, 3.99, 4.95, 5.25 for drives 1-5, respectively). See Figure 6 for graph. This main effect of drive supported Experiment 1 and 2 demonstrating that there are changes in trust at different points in time. To investigate the differences across the drives, a trend analysis was used to analyze the overall pattern of trust using within-groups contrasts. The analysis revealed there was not a significant quadratic trend, $F(1, 68) = 2.20$, $p = .143$, $\eta_p^2 = .03$. However, there was not a significant linear or cubic trendline either, *F*s < 1, indicating that no trend line could be fit to the data. Further analysis looking at the pairwise comparisons between drive 3 and the other drives showed that drive 3 was significantly different than all four other drives (*p*s < .001 for all). This analysis demonstrates the significant decrease in trust for the third drive compared to the initial and subsequent errorless drives. The pairwise comparisons provided support for the results found in Experiment 2, demonstrating that all conditions decreased during the error in drive 3. However, this did not support H3 and actually demonstrated the opposite of what was expected showing that active promotion repair did not positively influence trust and active dampening did not further decrease trust. There was a significant effect of the covariate ADOS, indicating that an individual's prior opinions on automated vehicles influenced their trust during the study, $F(1, 68) = 13.75$, $p < .001$, $\eta_p^2 = .17$. Using ADOS as a covariate, the analysis was able to control for initial levels of trust and opinions of automated driving, eliminating preconceived ideas about automation as a factor affecting the data.

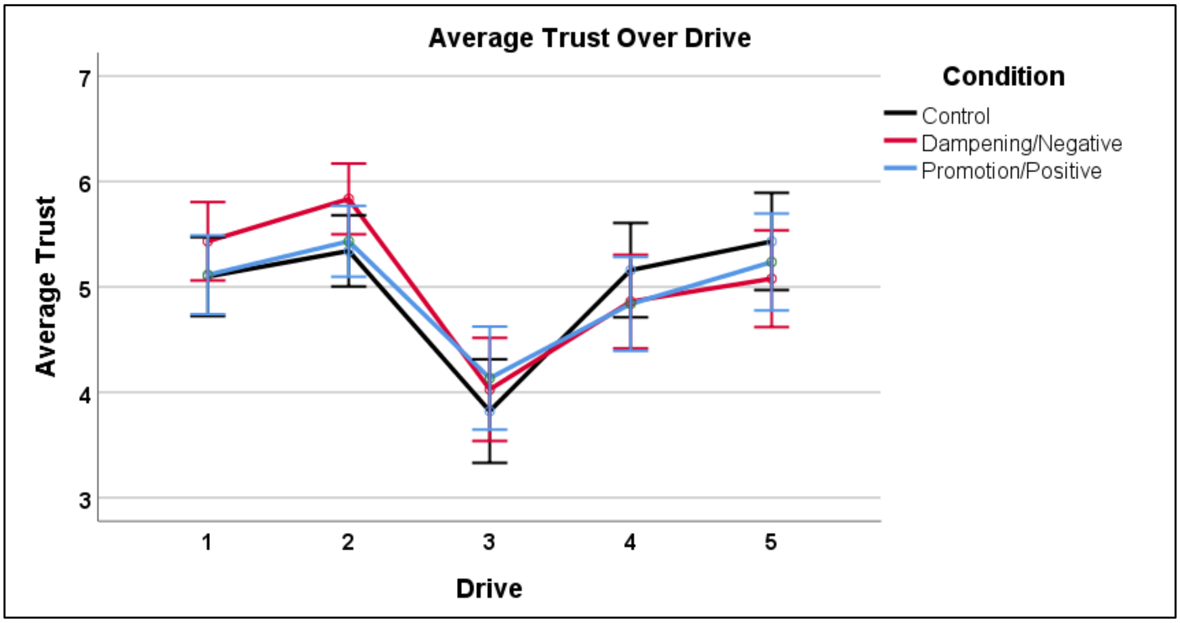
There was not a significant main effect of framing, $F(2, 68) = .09$, $p = .913$, $\eta_p^2 = .01$, showing that participants did not have a difference in trust after any framing. H1-H4 were not supported demonstrating that not only did the initial framings not replicate from the prior two experiments, but the active repair and active dampening strategies had not effect.

There was a significant interaction between drive and framing for trust, $F(8, 272) = 2.11$, $p = .035$, $\eta_p^2 = .06$. Simple main effects analysis showed only one significant difference control and dampening conditions in drive 2, $p = .043$. This did not match with any expected hypotheses because the framing conditions here seem to be leading the effect as the dampening and control conditions are switching after drive 3. I expected promotion to start higher, stay higher after the error on drive 3 and increase higher on drives 4 and 5, but this result was not shown. Though visually there seems to be variation in the drives, the results did not show much significant difference between conditions over drive.

H5 proposed an analysis integrating Experiment 2 and Experiment 3 as a quasi-experiment; however, due to the present findings, the analysis was disregarded as it is no longer necessary. However, I ran a new a quasi-experimental analysis to demonstrate that there were no major differences in the levels of trust for each drive. A t-test comparing the means of trust for each drive between Experiment 2 and Experiment 3 was performed. The tests showed that there was no significant difference in mean trust for any drive between Experiments 2 and 3, *p*s = .686, .994, .397, .426, and .497, for drives 1-5, respectively. This provides partial support for the idea that the groups did not have different interactions with the system and that trust repair strategies were overall not effective as they did not significantly shift trust.

Figure 6

Average trust over the five drives for each of the three conditions in Experiment 3



Note. The ADOS covariate appearing in the model was evaluated at 4.83 and the error bars are 95% Confidence Intervals.

2.3.3.2 Post-task Interview Results. This section contains the qualitative analyses from the post-experiment interview. The interview data was analyzed similarly to the data from experiment 2.

Participants were asked if they noticed anything strange about the automated driving system during their experience and a vast majority (95.83% or 69/72) of them mentioned the error of driving off the road either during this question, or at some point prior to the question being asked. The remaining participants that did not mention the error during this question or prior to the question were shown a short clip of the error and asked if they had seen this occur and all did see it and through later questioning did consider it an error of the automation. They had either forgotten to mention it or described it vaguely in previous statements that were not fully clarified.

Participants echoed very similar thoughts to those in Experiment 2 regarding why they thought the error occurred; however, those in the negative framing condition expressed more blame for the road having irregular road lines as the cause of the error because the trust dampening information provided this explanation. However, those in the control and positive framing condition had similar responses to participants in Experiment 2 because they were not told any specific cause of the error.

Again, participants echoed similar responses as Experiment 2 for the question about how the error influenced their trust. Two later questions asked specifically about the active calibration message (Q10, “What did you think about the message from the automated driving system after the drive where the error occurred?”) and if that influenced their trust (Q11, “How do you feel that this message influence your trust in the system?”). Participants in the control condition that saw a generic message about donating blood to the red cross stated that they did not associate the message with the error and didn’t think anything of it or thought that the message was strange and were uncertain why it showed up. A couple participants added that they liked it as a nice reminder and thought it was safe because they did not have to look at their phone.

For the positive and negative active calibration messages, responses were mixed with some participants liking the explanation and thinking that the message was potentially helpful to hear, compared to nothing. Others thought that if it could recognize the problem, they weren’t sure why it would not just avoid it or notify them during the drive. There were not many distinguishing themes between the positive and negative messages, perhaps indicating that the content of the messages could have been too similar resulting in no difference in effects. For over half participants in the positive (18/24) and negative (13/24) message conditions (31/48), the opinion was mixed with participants glad that it at least said something but were still concerned that it did not avoid the error. A few participants also mentioned that they wished the alert would have shown up sooner or wished it had shown up while the error was occurring. Similarly with the effect of the message on trust, those who had a mixed reaction also stated a mixed feeling of trust. There was no conclusive positive or negative impact of trust stemming from the message.

Again, similar to Experiment 2, participants had a diverse reaction when asked (Q4), “If you had the chance in real life, would you use the automated vehicle that would perform the same as the one you just saw in the study?” There were 44.44% of participants who said yes, 34.72% of participants who said no, and 27.78% of participants who were uncertain or needed more experience with the system first. Participants similarly expressed interest in checking it out but were still not fully certain about use all the time or in all situations.

2.3.4 Discussion

Experiment 3 had a few differences in results from the prior studies that did not turn out as expected. The active repair and active dampening strategies post-error did not have any different effects as we expected them to influence trust. It is hard to say whether this stemmed from the lack of replication from the previous studies or if it was due to a difficulty in actively repairing or dampening trust. The latter is very likely, considering that in much of trust repair research, there are a myriad of factors that can influence repair potential and this trend has been noted various times for individual differences such as working memory (Ku & Pak, 2023), differences in attention allocation (Sato et al., 2023) and for different types of automation systems or errors (Esterwood et al., 2023; Schelble et al., 2022; Xu & Howard, 2022), demonstrating that trust repair research is mutable and often many repair strategies are ineffective. An insight from the interview data is that participants who received the trust dampening information after the error stated more specific information about the potential cause of the error because they were told that was the reason. This increase in transparency or explanation might have influenced their trust; however, there were no differences between conditions so even through understanding a potential reason for the error, participants did not adapt their trust.

The obvious difference for Experiment 3 is that the trust promotion difference for the initial framing did not replicate. The effect of trust calibration framing did not appear the same as Experiments 1 and 2, and instead showed that trust did not differ across all conditions. Though the trend over the five drives was across conditions, the framing condition did not show the same effect. While this was unexpected, there are several viable explanations. First, I already discussed some limitations of trust repair, that it is fickle, depending on individual differences, mental models, and automation differences. Various underlying factors could impact individual's trust as a dynamic process and further as a sociocognitive process. Depending on an individual's relational understanding of an automated system at a current time, the framing could be more or less effective. Enacting swift trust could depend on more factors than reinforcing role and importing information. The interview data from this study highlights how certain individual differences can alter how people might respond differently to various situations. An individual's mental model about how a system works and behaves can influence their subsequent trust development process. Some individuals might view some mistakes as unforgiveable or have difficulty believing that an error can be corrected, requiring further proof of reliability over time. However, some individuals are more willing to trust – have higher dispositional trust – and their trust might be more malleable or easier to influence.

This leads to another potential issue: the specific wording of some of the frames were adapted for Experiment 3. The intent was to further align them as balanced opposites, but some positive or negative aspects may have been lost, which were unable to activate a change in trust. The wording of the positive framing was not changed but the control framing had additional neutral text added ("we have completed the installation of the update, and you may proceed with your drive now") and the negative framing changed from telling them they might need to watch to expressing that the reliability is uncertain ("you might need to keep watch to ensure everything goes well" to "we are not yet sure that it will perform perfectly"). The negative framing could be viewed as giving more transparency to the user about the reliability of the system, or that the idea of their role is less clear because they are not explicitly told that they might need to keep watch. These changes might have adjusted the base level trust of both control and negative framing, making them more equal to the unchanged positive framing, but this is unlikely. There is not enough evidence at this point to say that a wording change would be the cause of this difference, but it is worth highlighting.

One finding that Experiment 3 showed consistent with prior experiments was the significant differences across drives. The trend for the natural repair of trust – to increase after experience, decrease for an error, and then recover after the error – was shown, again backing up findings from Experiment 1 and 2, as well as prior research showing this similar trend (de Visser et al., 2018; Kraus et al., 2019; Mishler & Chen, 2023) but the active repair methods had no influence on this trust recovery. The trust framework through interacting with an automated driving system seems to consistently follow this trend and continual studies back it up. However, Experiment 3 was expected to displace this trend through the addition of active trust calibration after an error. I expected that providing an apology about the error with a promise to do better in the future would be able to lessen the decrease of trust. Prior research into automation errors has shown different results from a human-automation trust perspective depending on the type of error. This may be a running issue with trust repair research: errors are substantial and making a believable apology from an automation is difficult. The phenomenon of automation bias shows that errors from automated systems are more costly compared to their human counterparts (Dzindolet et al., 2003; Goodyear et al., 2017). The lack of robust findings for trust repair shows one way in which human-automation trust might be different from human-human trust. Humans are less likely to forgive automation errors and this could be due to the lack of belief that an automation can actually change. Often apologies or promises are offered as a band-aid without any actual change, and the human user might see through the veil if there are only words and not a tangible adjustment. I thought that by adding apology – one of the most widely applied trust repair methods – with promise, it might benefit trust repair, as combinations can be helpful (de

Visser et al., 2018) and some work on using promises as repair have been marginally successful, albeit not always for overall trust measures (Esterwood et al., 2023; Xu & Howard, 2022).

Another explanation for potential lack of findings for trust repair is that the apology took place too late. After the error occurred, the vehicle fully recovered and spent up to thirty seconds finishing the drive before the end, where the message was displayed. After the error, the vehicle drove on a straight section of the road with no further incidents, but only maintained speed and stayed in the lines. I expected that the apology – and the dampening – would still be given in sufficient time letting them settle with the situation and then process the repair or dampening intervention. The idea that some trust repair or recovery might have already happened can be seen from the interview data, showing that participants seeing good performance after the error started to reinstate their trust in the system. Additionally, several participants mentioned that they wish the message had happened sooner or while the error occurred and not so late after. Other research has shown that timing can have strong influence over the effectiveness of trust repair strategies. Kohn and colleagues (2018) found timed apology to be the most effective strategy, and Robinette and colleagues (2015) along with Nayyar and Wagner (2018) found that repair strategies during an emergency and early soon after were more effective for an office evacuation with robot interaction task. For Experiment 3, I did not want to display a visual warning during the drive that would take participants' eyes off the screen during the drive right after the error in case they were still concerned that another error might occur. If the text was presented right after the error, they might not have even trusted the system enough to continue driving unmonitored while they read the message. However, other methods for displaying the message sooner and in a way that is easier to process and understand could be considered for more effective trust repair.

CHAPTER 3

GENERAL DISCUSSION AND CONTRIBUTIONS

The findings of this study contributed to the theoretical understandings of trust and trust calibration towards newly introduced automation. As the use of automation increases, humans will be interacting and teaming with new automated technology in numerous parts of their lives. Trust in automation has major implications for the use, misuse, or disuse of automation, so understanding factors that affect trust can have far reaching-consequences if they are not accounted for (Parasuraman & Riley, 1997). Interactions with new and recently updated technology can affect adoption of a technology or how much user oversight is given to the functions during its use. Some individuals might never use a system if they do not trust it, losing out on any helpful benefits it can offer. Conversely, some individuals might misuse a system if trust it too much, allowing it far more unmoderated action than it should have. This study contributes to the current literature on trust calibration by showing how initial framing, errors, and active calibration strategies influence trust.

This study implemented swift trust to contribute to the understanding of how human-automation teams quickly develop trust to work together after new automation updates have been implemented. Well known framing literature was applied in a unique way to inspire initial levels of trust for the new automation features to uniquely tune the humans' trust calibration. Positive and negative framing has been shown to alter risk taking behavior. The current study has expanded those findings to framing an individual's initial swift trust through highlighting positive and negative aspects of an automated system. This is beneficial for introducing new human-automation teams and helping to calibrate their trust. The swift trust literature for human-human teams has tried to increase imported trust at initial meetings through active engagement strategies meant to boost team trust with limited benefit (Kroeger et al. 2021). Initial human interactions can be highly variable, depending on context and cues that cannot be controlled (Haring et al., 2021). The current study showed that it is possible to create different trust outcomes depending on the framing of initial swift trust, contributing to the scarce literature on applying swift trust to human-automation teams. It showed the importance of how initial framing of the automation's role and capabilities influences human trust in automation which has important implications for overtrust and undertrust in later interactions.

The outcomes for implementing framing of swift trust in this study demonstrate the importance of instruction and prior information before individuals interact with an automated system. The capacity to promote or dampen trust before someone interacts with a system can be greatly beneficial in helping to properly match their trust calibration with the performance of the system. However, trust calibration is more complex than just pairing trust and performance. The performance of an ADS can be highly variable depending on the level of automation, driving and road conditions, brand of the vehicle, and many more variables. It is extremely challenging to match a user's expected trust in the system with the performance of a system and therefore chasing a perfect trust calibration is worthless. Instead, it is important to consider what level of calibration would be acceptable and try to fit the user into that range while keeping in mind the acceptable reliability of the system. An acceptable level of calibration would be one that results in the highest level of safety. Trust in automation is constantly adapting and changing, especially among individuals. This study showed that participants' trust in a system could be influenced by changing the way of describing an automated update, and any subsequent overtrust in a system like an ADS could end up costing someone's life. On the other hand, many systems are performing better than a human and in a complex environment like driving, human drivers cannot account for everything. If an individual thinks they can manage without a safety system or tries to override a system that could be preventing a crash, this again could result in bodily harm or loss of life. Disregard for the human being's understanding of the automated system while placing them in a dangerous situation is tantamount to condemning them to failure. In a new system that users are unfamiliar with, it is too easy to alter their trust. Designers and researchers should be wary of human drivers using features that they do not understand which have the potential for harm.

In addition to manipulating swift trust calibration at the beginning of the interaction, I tried to actively calibrate trust after interaction with the system and the experience of a system error. To keep in line with trust promotion or trust dampening, strategies either intended to nullify the negative effects of the error on trust (positive/promotion) or highlight the error and explain the weaknesses of the system to temper trust (negative/dampening) were implemented to further understand how the automation in human-automation teams can influence trust. Initial influences on trust are important, but the longer lasting implications of these influences might be affected by the team interaction over time. Understanding how the trust calibration is influenced through interaction is important for these novel situations. Automation errors are bound to happen as no automation is perfect. Human-human teams often implement repair strategies or social responses after trust violations, whether as a premeditated plan or as an innate reaction to doing something wrong (de Visser et al., 2020; Madhavan & Wiegmann, 2007). Testing active strategies of trust calibration after an error is still relatively unexplored in the field of human-automation interaction, especially for autonomous vehicles. For a new automation system or new update, designers might need to dampen trust to prevent users from misusing the system if the automation reliability is not fully verified. Alternatively, certain situations might call for trust promotion when an error can be learned from and actively corrected for the future. Automation systems might be able to dynamically implement these different strategies depending on what is required, and this study provides insights into the entire process of framing initial trust and trying to calibrate the trust over time.

The results of the study show that active trust calibration is not simple. Some participants were very impressed by the automation and its ability to detect the error and say that it would correct it, while others doubted the validity of these messages and simply refused to trust the system. Most participants were somewhere in the middle, left uncertain about the future performance of the system. Designers need to be careful how they communicate the information about the system to the end user and what the user is told during the interaction. The metrics and capacities of the system are not always known or easily understood by the end user, so the way in which they are told what is happening or what they should be expecting can be highly impactful on their experience and future actions.

If there is a lot of text or documentation, one thing that can almost be guaranteed is that people will not read it all or might not read any of it. Therefore, according to best practices, the wording and content was made to be as short and simple as possible (Brust-Renck et al., 2013; Lacson et al., 2005; Norman, 1983). By making the text short it allowed users to quickly intake the message. However, there are many other facets to the information that could be displayed. The focus on role and importing trust were thought to be the most important, but there could be other categories of information that could be important. Balancing detail and information length is complicated, but enough information was provided to get a positive response for this study.

The promotion and dampening information from the current study was abridged and did not always fully explain exactly what was happening with the vehicle, meaning that the system transparency was insufficient in this case. The information was given to them after the fact and not as the error was unfolding, which could have contributed to a lesser effect of the repair effect. This is one of the core aspects of transparency that it should show the steps to complete a successful task, or the specific data being used at a point in time to make a decision. The vehicle should show the transparency of the system at the time to allow the human user to understand what is happening. In contrast to the system notifying the user after something has occurred and relying on the user remembering the error and recoding their experience of the memory to update what might have been going on with the system. While the error was occurring, the participant was not able to have control and therefore might not have been able to repair trust due to the lack of empowerment of the user. However, when providing transparency information, one needs to ensure that the user knows what to do with that information. Increasing transparency can increase an individual's use of the system without increasing their understanding. The individual might have a surface-level way of thinking about the information, explanation, or response that they have not generated through their own thought. The user might simply accept the information from the system, even if that is not optimal. Some potential explanations of an ADS's behavior used for trust repair or dampening might not be understood or processed the same way by different individuals, especially if they do not have a proper mental model or context for understanding what happened. Therefore, careful thought should go into the design of transparency or explainable behavior from the system to the user if the user might take away the wrong message.

In Experiment 2, trust promotion was shown to increase trust, and the effect even lasted after an error. This finding is of great importance relative to calibration because it shows a potential for miscalibration of trust. Numerous studies have demonstrated an overtrust for miss-prone automation (Bailey & Scerbo, 2007; de Visser et al., 2016; Dzindolet et al., 2001; Mishler et al., 2019; Molloy & Parasuraman, 1996). However, many studies or designers also recommend ways to improve trust in automation through design alterations not related to a system’s actual reliability, but instead its design features or preexisting knowledge such as anthropomorphism, appearance, and communication style, brand reputation (de Visser et al., 2016; Endsley, 2017; Hoff & Bashir, 2015; Sanchez et al., 2014). To be clear, many of these studies are not simply recommending increasing trust as much as possible and do consider a proper calibration of trust to be imperative. The closest designers can come to proper trust calibration is one that allows the user to let the automation function without intervention for as long as reasonable, but still be ready and capable of intervening when necessary.

Experiments 1 and 2 demonstrated that trust calibration is fickle. Even though drivers experience the same exact ADS, their trust is uniquely calibrated simply by a manipulation to swift trust before interaction with the system. Experiment 2 showed the potential downfall of simply designing to inspire automation trust because even though the system drove off the road, users that were primed to have increased trust had higher trust recovery after that error compared to the other conditions, demonstrating an easily manipulable ‘bias’ that can be introduced. This directly falls in line with Endsley’s (2017) study, which found that sellers of a new Tesla vehicle misrepresented the capabilities of the system and oversold its ability to drive and spot hazards. If a user did not know the system’s limitations and the capabilities were overexaggerated, the driver might not be able to properly calibrate their trust even after interaction with the system. Interaction with the system and dynamically learned trust is one of the highly touted ways that individuals learn about a system establish a proper calibration, but if this initial learned trust can be manipulated beforehand, it could fundamentally alter their later interactions with the system leading to problematic and dangerous behavior.

Overall, the results show that trust can be easily influenced, leading to a more in depth understanding of trust as a physiological construct. Due to the nature of human cognition, decisions can be quickly influenced, and new information can alter risk taking behavior. This study has demonstrated that the same can be said about trust in automation. Trust is a behavioral reaction to an understanding of a system, and perceived opinion can be altered. In contrast to traditional prospect theory expecting that shows loss aversion, my study showed greater benefits of gain (Kahneman & Tversky, 1979). Typically, risk incites greater emotional impact; however, the gain/promotion condition of Experiment 2 only showed an increase in trust after the error, demonstrating the swift trust building from the initial framing. Though Experiment 1 showed increases for promotion and decreases for dampening, a further difference was uncovered when trust was tested during the error in Experiment 2. Trust is understood to be a variety of preexisting knowledge and initial learned trust prior to interacting with a system (Hoff & Bashir, 2015), but this study demonstrated that initial learned trust could be manipulated reflecting a change in trust. Trust is also understood to be variable in nature; however, the variability of trust is often said to be developed during interaction in the dynamic learning of trust. The present study demonstrates the variability in trust calibration outcomes through initial information given before interaction. Trust does develop during repeated interaction, but the path of development during interaction can be altered by adjustments just prior to the interaction.

This study only investigated the occurrence of one error; however, in many situations there might be repeated errors which could change their trust calibration. The difficulty of the driving situation and the severity of such errors could also impact drivers' trust. The natural recovery of trust after an error found in this study and prior research (Kraus et al., 2019; Mishler & Chen, 2023) could be lesser after a more severe error or after multiple errors either in quick succession or spread over time. Additionally, because of the increased trust recovery for the positive framing, it is unclear if the framing would still be as effective after multiple errors. After a certain number of errors, disuse becomes a more and more likely outcome. The positive framing could influence the users to continue use of the system past the point of other drivers as errors stack up. Potentially, positive framing could make trust more resilient to errors over time and error severity though further research is needed to uncover further impacts of the early initial trust development process.

Something not always discussed in the trust in automation literature is the ethics of design. Often people discuss the ‘trolley problem’ in relation to how an automated vehicle might choose to kill a pedestrian to save the human occupant or vice-versa (Bruers et al., 2014; Wu, 2020). The trolley problem works from a theoretical standpoint and makes for an interesting thought experiment. But no automotive company ought to state that they are making such a binary choice for such a complex situation, though some are trying to create risk cost functions for trajectory planning (Geisslinger et al., 2021). However, the design features and system introduction/training are a more practical ethical quandary because improper design or training can and has led a user to misuse a two-ton system of death.

A potential issue regarding combining trust promotion and then having participants see an error after a relatively short time is the contrast between assuring them that the automation would be performing well (promotion) and then seeing an error. This resulting contrast effect could have erased any effect of trust promotion and even had the inverse effect after an error, further decreasing trust. Whereas with dampening trust, participants might have been expecting an error, then when an error occurred, it confirmed their suspicions. It would be different to determine the latter in this study because there was no difference found for the trust dampening and control conditions for Experiments 2 and 3. There was not a contrast effect in this experiment for promotion because trust did not decrease different from the other conditions during the error, and still recovered higher after the error.

One limitation with the current study is that trust is a hard concept to study, especially though surveys. Many different types of surveys are used in trust research, and there is not a general consensus on the best to use, and none have been solidly established for trust in ADS, though many are in development. The current trust survey used (Jian et al., 2000) is widely accepted and validated; however, it might not be able to fully capture the full range of positive and negative changes in trust. The survey was not developed with the thought of continual repeated use to measure trust changes over time or developed for newer frameworks that consider the full aspects of dispositional, situational, and learned trust influencing interaction (Chiou & Lee, 2021; Hoff & Bashir, 2015). Another limitation is the potential that the framing text might not have been equivalently balanced in terms of content and information, even though there was an attempt to balance them. There can be a lot of different combinations of information given to participants or the phrasing to promote certain aspects could be adjusted and might not be equivalent across groups of individuals. In future studies, deeper examination into various different framing information to influence swift trust could be explored to see if some are more effective at eliciting higher and lower trust. A final limitation is that this was all run on a simulator. Though simulators can of sufficient fidelity, some of the risk involved might not be conveyed to the participant as strongly as in a real automated vehicle. If the study was done using a real system on the roads, the added risk could produce larger effects.

CHAPTER 4

CONCLUSION

Overall, the findings of this study helped explain the trust process for new human-automation teams, leading to a clearer picture of the teamwork dynamic between this emerging enterprise. New automation systems with various new updates are ubiquitous in modern life and understanding how to get users to team with and work alongside automation is a necessity. This study provided some of the first applications of trust-calibration manipulation both before and during interaction for autonomous vehicles and following an automation error. These results demonstrate a unique understanding for trust in automation as it relates to development before interaction. The framing of the automation before interaction has lasting effects for trust development and calibration. This contributes to further understandings of initial trust before interaction and how the impact of positive framing further demonstrates the variability of trust. The idea that trust is developed, calibrated, and established during interaction is incomplete, and there is much more research to be done in the factors that influence trust based on initial information, mental models, and automation understanding before interaction. In terms of design implications, future designers should look to implement trust calibration strategies to ensure automation and humans can work together as a team. Determining the proper way to explain the system, its performance, capabilities, and role are all important for helping the user to properly calibrate their trust. Knowing how much one can rely on a teammate or when they need to assist or monitor another teammates behavior can become critically important for successful teams. Additionally, the findings highlighted many potential individual differences that are not yet fully explored. A consideration for future designers and researchers is to consider the audience you are designing for and understand their mental models of the system beforehand. Overall, this study lays out a foundation for future trust calibration research and design between human-automation teams.

REFERENCES

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020). Trust repair in performance, process, and purpose factors of human-robot trust. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–6.

Annett, J., & Stanton, N. (2006). Task analysis. *International Review of Industrial and Organizational Psychology*, *21*, 45.

Azevedo-Sa, H., Jayaraman, S. K., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2020). Context-Adaptive Management of Drivers' Trust in Automated Vehicles. *IEEE Robotics and Automation Letters*, *5*(4), 6908–6915. https://doi.org/10.1109/LRA.2020.3025736

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. https://doi.org/10.1080/14639220500535301

Baker, A., Phillips, E., Ullman, D., & Keebler, J. (2018). Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), 1–30. https://doi.org/10.1145/3292532

Bakker, R. M. (2010). Taking Stock of Temporary Organizational Forms: A Systematic Review and Research Agenda. *International Journal of Management Reviews*, *12*(4), 466–486. https://doi.org/10.1111/j.1468-2370.2010.00281.x

Balfe, N., Sharples, S., & Wilson, J. R. (2018). Understanding Is Key: An Analysis of Factors Pertaining to Trust in a Real-World Automation System. *Human Factors*. https://doi.org/10.1177/0018720818761256

Banks, V. A., Eriksson, A., O'Donoghue, J., & Stanton, N. A. (2018). Is partially automated driving a bad idea? Observations from an on-road study. *Applied Ergonomics*, *68*(August 2017), 138–145. https://doi.org/10.1016/j.apergo.2017.11.010

Bass, B., Goodwin, M., Brennan, K., Pak, R., & McLaughlin, A. (2013). Effects of age and gender stereotypes on trust in an anthropomorphic decision aid. *Proceedings of the Human Factors and Ergonomics Society*, *September*, 1575–1579. https://doi.org/10.1177/1541931213571351

Benamati, J., Serva, M. A., & Fuller, M. A. (2006). Are trust and distrust distinct constructs? An empirical study of the effects of trust and distrust among online banking users. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, *6*, 121b–121b.

Bennett, J. M., Challinor, K. L., Modesto, O., & Prabhakharan, P. (2020). Attribution of blame of crash causation across varying levels of vehicle automation. *Safety Science*, *132*(June), 104968. https://doi.org/10.1016/j.ssci.2020.104968

Benson, A. J., Tefft, B. C., Svancara, A. M., & Horrey, W. J. (2018). Potential reductions in crashes, injuries, and deaths from large-scale deployment of advanced driver assistance systems. *Research Brief*.

Bethel, C. L., Salomon, K., Murphy, R. R., & Burke, J. L. (2007). Survey of psychophysiology measurements applied to human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 732–737. https://doi.org/10.1109/ROMAN.2007.4415182

Brust-Renck, P. G., Royer, C. E., & Reyna, V. F. (2013). Communicating Numerical Risk: Human Factors That Aid Understanding in Health Care. In *Reviews of Human Factors and Ergonomics* (Vol. 8, Issue 1, pp. 235–276). https://doi.org/10.1177/1557234X13492980

Capiola, A., Baxter, H. C., Pfahler, M. D., Calhoun, C. S., & Bobko, P. (2020). Swift Trust in Ad Hoc Teams: A Cognitive Task Analysis of Intelligence Operators in Multi-Domain Command and Control Contexts. *Journal of Cognitive Engineering and Decision Making*, *14*(3), 218–241. https://doi.org/10.1177/1555343420943460

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors*, *59*(3), 333–345. https://doi.org/10.1177/0018720816682648

Chen, J., Gates, C. S., Li, N., & Proctor, R. W. (2015). Influence of risk/safety information framing on android app-installation decisions. *Journal of Cognitive Engineering and Decision Making*, *9*(2), 149–168. https://doi.org/10.1177/1555343415570055

Chen, J., Mishler, S., & Hu, B. (2021). Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems*, *51*(5), 463–473. https://doi.org/10.1109/THMS.2021.3051137

Chen, J., Mishler, S., Hu, B., Li, N., & Proctor, R. W. (2018). The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *International Journal of Human Computer Studies*, *119*(November 2017), 35–47. https://doi.org/10.1016/j.ijhcs.2018.05.010

Chen, J., Mishler, S., Long, S., Yahoodik, S., Garcia, K., & Yamani, Y. (2022). Human-Automation Interaction for Semi-Autonomous Driving: Risk Communication and Trust. In *Human-Automation Interaction: Transportation* (pp. 281–291). Springer.

Chiou, E. K., & Lee, J. D. (2021). Trusting Automation: Designing for Responsivity and Resilience. In *HUMAN FACTORS* (Vol. 00, Issue 0).

Cicchino, J. B. (2017). Effectiveness of forward collision warning and autonomous emergency braking systems in reducing front-to-rear crash rates. *Accident Analysis & Prevention*, *99*, 142–152.

Coppola, N. W., Hiltz, S. R., & Rotter, N. G. (2004). Building trust in virtual teams. *IEEE Transactions on Professional Communication*, *47*(2), 95–104. https://doi.org/10.1109/TPC.2004.828203

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331–349. https://doi.org/10.1037/xap0000092

de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 263–267.

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From ‘automation’ to ‘autonomy’: the importance of trust repair in human–machine interaction. *Ergonomics*, 1–19. https://doi.org/10.1080/00140139.2018.1457725 de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

de Visser, E., & Parasuraman, R. (2011). Adaptive Aiding of Human-Robot Teaming:Effects of Imperfect Automation on Performance, Trust, and Workload. *Journal of Cognitive Engineering and Decision Making*, *5*(2), 209–231. https://doi.org/10.1177/1555343411410160

Degani, A., Shafto, M., & Kirlik, A. (1999). Modes in human-machine systems: Constructs, representation, and classification. *International Journal of Aviation Psychology*, *9*(2), 125–138. https://doi.org/10.1207/s15327108ijap0902_3

DeGuzman, C. A., & Donmez, B. (2021a). Drivers still have limited knowledge about adaptive cruise control even when they own the system. *Transportation Research Record*, *2675*(10), 328–339.

DeGuzman, C. A., & Donmez, B. (2021b). Knowledge of and trust in advanced driver assistance systems. *Accident Analysis & Prevention*, *156*, 106121.

DeGuzman, C. A., Hopkins, S. A., & Donmez, B. (2020). Driver Takeover Performance and Monitoring Behavior with Driving Automation at System-Limit versus System-Malfunction Failures. *Transportation Research Record*, *2674*(4), 140–151. https://doi.org/10.1177/0361198120912228

Dekker, S. W. A., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human–automation co-ordination. *Cognition, Technology & Work*, *4*, 240–244.

Demir, M., McNeese, N. J., Gorman, J. C., Cooke, N. J., Myers, C. W., & Grimm, D. A. (2021). Exploration of Teammate Trust and Interaction Dynamics in Human-Autonomy Teaming.

IEEE Transactions on Human-Machine Systems, *51*(6), 696–705. https://doi.org/10.1109/THMS.2021.3115058

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Du, N., Huang, K. Y., & Yang, X. J. (2020). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, *62*(6), 987–1001. https://doi.org/10.1177/0018720819862916

Dunn, N. J., Dingus, T. A., Soccolich, S., & Horrey, W. J. (2021). Investigating the impact of driving automation systems on distracted driving behaviors. *Accident Analysis & Prevention*, *156*, 106152.

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

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Esterwood, C., & Robert, L. P. (2021). Do you still trust me? human-robot trust repair strategies. *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, 183–188.

Esterwood, C., Ali, A., George, Z., Dubrow, S., Smereka, J., Riegner, K., Tilbury, D., & Jr, L. P. R. (2023). Promises and Trust Repair in UGVs. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/21695067231196235

Fallon, C. K., Bustamante, E. A., Ely, K. M., & Bliss, J. P. (2005). Improving user trust with a likelihood alarm display. *Proceedings of the 1st Conference on Augmented Cognition, Las Vegas, NV*.

Ford Motor Company. (2016). *Ford Targets Fully Autonomous Vehicle for Ride Sharing in 2021; Invests in New Tech Companies, Doubles Silicon Valley Team*. https://media.ford.com/content/fordmedia/fna/us/en/news/2016/08/16/ford-targets-fully-autonomous-vehicle-for-ride-sharing-in-2021.html

Fox, J. E., & Boehm-Davis, D. A. (1998). Effects of age and congestion information accuracy of advanced traveler information systems on user trust and compliance. *Transportation Research Record*, *1621*(1), 43–49.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in Automation – Before and After the Experience of Take-over Scenarios in a Highly Automated Vehicle. *Procedia Manufacturing*, *3*(Ahfe), 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847

Goodyear, K., Parasuraman, R., Chernyak, S., de Visser, E., Madhavan, P., Deshpande, G., & Krueger, F. (2017). An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social Neuroscience*, *12*(5), 570–581. https://doi.org/10.1080/17470919.2016.1205131

Greenberg, P. S., Greenberg, R. H., & Antonucci, Y. L. (2007). Creating and sustaining trust in virtual teams. *Business Horizons*, *50*(4), 325–333. https://doi.org/10.1016/j.bushor.2007.02.005

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*(2), 284–291. https://doi.org/10.1080/00140139.2016.1190035

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*, 517–527. https://doi.org/10.1177/0018720811417254

Haring, K. S., Phillips, E., Lazzara, E. H., Ullman, D., Baker, A. L., & Keebler, J. R. (2021). Applying the swift trust model to human-robot teaming. In *Trust in Human-Robot Interaction* (pp. 407–427). Elsevier.

Hawkins, A. J. (2018, June 19). *'Autopilot Buddy' that tricks Tesla vehicles declared 'unsafe' by US*. The Verge. https://www.theverge.com/2018/6/19/17479316/tesla-autopilot-buddy-aftermarket-nhtsa

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior Familiarization with Takeover Requests Affects Drivers' Takeover Performance and Automation Trust. *Human Factors*, *59*(3), 457–470. https://doi.org/10.1177/0018720816678714

Hergeth, S., Lorenz, L., Krems, J. F., & Toenert, L. (2015). *Effects of Take-Over Requests and Cultural Background on Automation Trust in Highly Automated Driving*. 331–337. https://doi.org/10.17077/drivingassessment.1591

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hong, J. W., & Curran, N. M. (2019). Artificial intelligence, artists, and art: Attitudes toward artwork produced by humans vs. artificial intelligence. *ACM Transactions on Multimedia Computing, Communications and Applications*, *15*(2s). https://doi.org/10.1145/3326337

International Organization for Standardization. (2023). *Road vehicles — Software update engineering (ISO Standard No. 24089)*. https://www.iso.org/standard/77796.html

Jamieson, G. A., & Skraaning, G. (2018). Levels of Automation in Human Factors Models for Automation Design: Why We Might Consider Throwing the Baby Out With the Bathwater. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 42–49. https://doi.org/10.1177/1555343417732856

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/s15327566ijce0401_04

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 7–24. https://doi.org/10.1177/1555343417737203

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10.2307/1914185

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice*, *94*, 182–193. https://doi.org/10.1016/j.tra.2016.09.010

Katrakazas, C., Quddus, M., Chen, W. H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C: Emerging Technologies*, *60*, 416–442. https://doi.org/10.1016/j.trc.2015.09.011

Kauffmann, N., Winkler, F., Naujoks, F., & Vollrath, M. (2018). "What Makes a Cooperative Driver?" Identifying parameters of implicit and explicit forms of communication in a lane change scenario. *Transportation Research Part F: Traffic Psychology and Behaviour*, *58,* 1031–1042.

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, *99*(1), 49–65.

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (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), 104.

Kohn, S. C., Quinn, D., Pak, R., de Visser, E. J., & Shaw, T. H. (2018). Trust Repair Strategies with Self-Driving Vehicles: An Exploratory Study. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, *62*(1), 1108–1112. https://doi.org/10.1177/1541931218621254

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors*. https://doi.org/10.1177/0018720819853686

Kroeger, F., Racko, G., & Burchell, B. (2021). How to create trust quickly: A comparative empirical investigation of the bases of swift trust. *Cambridge Journal of Economics*, *45*(1), 129–150. https://doi.org/10.1093/cje/beaa041

Ku, C., & Pak, R. (2023). The Effects of Individual Differences in Working Memory on Trust Recovery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. https://doi.org/10.1177/21695067231195000

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345–360. https://doi.org/10.1080/00140139.2018.1547842

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). Effects of Attribute and Goal Framing on Automation Reliance and Compliance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 482–486. https://doi.org/10.1177/154193120504900357

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, *8*(4), 355-362.

Langfred, C. W. (2004). Too Much of a Good Thing? Negative Effects of High Trust and Individual Autonomy in Self-Managing Teams. In *Source: The Academy of Management Journal* (Vol. 47, Issue 3). https://www.jstor.org/stable/20159588?seq=1&cid=pdf-

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics and Computing*, *13*(2), 163–167.

Lee, J. D. (2008). Fifty years of driving safety research. *Human Factors*, *50*(3), 521–528.

Lee, J. D. (2018). Perspectives on automotive automation and autonomy. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 53–57.

Lee, J. D., & See, K. a. (2004). Trust in automation: designing for appropriate reliance. *Human Factors*, *46*, 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015). A framework for rebuilding trust in social automation across health-care domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *4*(1), 201–205.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*. https://doi.org/10.1518/001872006779166334

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial Trust Formation in New Organizational Relationships. In *Source: The Academy of Management Review* (Vol. 23, Issue 3). https://www.jstor.org/stable/259290

McLaughlin, A. C., & Mayhorn, C. B. (2014). Designing effective risk communications for older adults. *Safety Science*, *61*, 59–65. https://doi.org/10.1016/j.ssci.2012.05.002

Metcalfe, J. S., Marathe, A. R., Haynes, B., Paul, V. J., Gremillion, G. M., Drnec, K., Atwater, C., Estepp, J. R., Lukos, J. R., Carter, E. C., & Nothwang, W. D. (2017). Building a framework to manage trust in automation. *Micro- and Nanotechnology Sensors, Systems, and Applications IX*, *10194*, 101941U. https://doi.org/10.1117/12.2264245

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196–204. https://doi.org/10.1518/hfes.46.2.196.37335

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

Mishler, S. (2019). Whose Drive Is It Anyway? Using Multiple Sequential Drives to Establish Patterns of Learned Trust , Error Cost , and Non-Active Trust Repair While Considering Daytime and Nighttime Differences as a Proxy for Difficulty. *(Master's Thesis)*, *Retrieved*. https://doi.org/10.25777/e3ce-8x78

Mishler, S., & Chen, J. (2023). Effect of automation failure type on trust development in driving automation systems. *Applied Ergonomics*, *106*. https://doi.org/10.1016/j.apergo.2022.103913

Mishler, S., Chen, J., Sabic, E., Hu, B., Li, N., & Proctor, R. W. (2017). Description-experience gap: The role of feedback and description in human trust in automation. *Proceedings of the Human Factors and Ergonomics Society*. https://doi.org/10.1177/1541931213601559

Mishler, S., Jeffcoat, C., & Chen, J. (2019). Effects of Anthropomorphic Phishing Detection Aids , Transparency Information , and Feedback on User Trust , Performance , and Aid Retention. *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting, 2000,* 2019. https://doi.org/10.1177/1071181319631351

Molloy, R., & Parasuraman, R. (1996). Monitoring an Automated System for a Single Failure: Vigilance and Task Complexity Effects. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*, 311–322. https://doi.org/10.1177/001872089606380211

Montague, E., Xu, J., & Chiou, E. (2014). Shared experiences of technology and trust: An experimental study of physiological compliance between active and passive users in technology-mediated collaborative encounters. *IEEE Transactions on Human-Machine Systems*, *44*(5), 614–624. https://doi.org/10.1109/THMS.2014.2325859

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man-Machine Studies*, *27*, 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*, 429–460. https://doi.org/10.1080/00140139608964474

National Transportation Safety Board. (2018). *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian*.

Nayyar, M., & Wagner, A. R. (2018). When should a robot apologize? understanding how timing affects human-robot trust repair. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11357 LNAI*, 265–274. https://doi.org/10.1007/978-3-030-05204-1_26

Norman, D. A. (1983). Design principles for human-computer interfaces. *Conference on Human Factors in Computing Systems - Proceedings*, *December*, 1–10. https://doi.org/10.1145/800045.801571

Overton, T. L., Rives, T. E., Hecht, C., Shafi, S., & Gandhi, R. R. (2015). Distracted driving: prevalence, problems, and prevention. In *International Journal of Injury Control and Safety Promotion* (Vol. 22, Issue 3, pp. 187–192). https://doi.org/10.1080/17457300.2013.879482

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*. https://doi.org/10.1518/001872097778543886

Read, G. J. M., Shorrock, S., Walker, G. H., & Salmon, P. M. (2021). State of science: evolving perspectives on 'human error.' In *Ergonomics* (Vol. 64, Issue 9, pp. 1091–1114). Taylor & Francis. https://doi.org/10.1080/00140139.2021.1953615

SAE. (2021). Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. In *SAE International* (Vol. J3016, pp. 1–41). https://doi.org/https://doi.org/10.4271/J3016_202104

Safar, J. A., & Turner, C. W. (2005). Validation of a two factor structure for system trust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*(3), 497–501.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160. https://doi.org/10.1080/1463922X.2011.611269

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*(1), 5–19. https://doi.org/10.1518/001872095779049516

Sato, T., Inman, J., Politowicz, M. S., Chancey, E. T., & Yamani, Y. (2023). *A Meta-Analytic Approach to Investigating the Relationship Between Human-Automation Trust and Attention Allocation*. 1–6. https://doi.org/10.1177/21695067231194333

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2022). Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors*. https://doi.org/10.1177/00187208221116952

Seong, Y., & Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, *38*(7–8), 608–625. https://doi.org/10.1016/j.ergon.2008.01.007

Singer, J., & Jenness, J. W. (2020). *Impact of information on consumer understanding of a partially automated driving system*.

Singh, S. (2018). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey (Traffic Safety Facts Crash•Stats. Report No. DOT HS 812 506). *National Highway Traffic Safety Administration*, *March*, 1–3.

Smiley, A. (2000). Auto safety and human adaptation. *Issues in Science and Technology*, *17*(2), 70–76.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008). Towards an empirically developed scale for system trust: Take two. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(19), 1335–1339.

Tesla. (2017). *Tesla Autonomous Mode Disengagements for Reporting year 2017*. *https://www.dmv.ca.gov/portal/wcm/connect/f965670d-6c03-46a9-9109-0c187adebbf2/Tesla.pdf?MOD=AJPERES*. https://www.dmv.ca.gov/portal/wcm/connect/f965670d-6c03-46a9-9109-0c187adebbf2/Tesla.pdf?MOD=AJPERES

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. In *Science* (Vol. 211, Issue 441, pp. 453–458).

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human Factors*, *60*(8), 1095–1116.

Wildman, J. L., Shuffler, M. L., Lazzara, E. H., Fiore, S. M., Burke, C. S., Salas, E., & Garven, S. (2012). Trust development in swift starting action teams: A multilevel framework. *Group and Organization Management*, *37*(2), 137–170. https://doi.org/10.1177/1059601111434202

Xu, J., & Howard, A. (2022). Evaluating the Impact of Emotional Apology on Human-Robot Trust. *RO-MAN 2022 - 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots*, 1655–1661. https://doi.org/10.1109/RO-MAN53752.2022.9900518

APPENDIX A

DEMOGRAPHIC INFORMATION SURVEY

					Gender		Ethnicity (please select one)							Dominant Hand	Normal or Corrected to Normal	
Date	RA	Expt	Subject No.	Age	M	F	American Indian/ Alaska Native	Asian	African/ African American	Caucasian	Native Hawaiian/ Pacific Islander	More than one race	Other/ Unknown	L/R	Vision (Y/N)	Hearing (Y/N)

APPENDIX B

AUTOMATED DRIVING OPINION SURVEY

The purpose of this survey is to gather your experience with manual driving and your opinions on automated driving. The following is a description of the levels of driving automation from the 2016 SAE Taxonomy and Definitions for Driving Automation systems that is referenced in the following survey.

- **Manual driving:** The human driver executes the driving task him/herself using the steering wheel and pedals.
- **Partially Driving Automation:** The automated driving system takes over both speed and steering control on some roads. However, the system cannot handle all possible situations. Therefore, the driver shall permanently monitor the road and be prepared to take over control at any time.
- **Conditional Driving Automation:** The automated driving system takes over both speed and steering control on most roads. The driver is not required to permanently monitor the road. If automation cannot handle a situation it provides a take-over request, and the driver must take-over control with a time buffer of 7 s.
- **Highly Driving Automation:** The automated driving system takes over both speed and steering control on all roads. The driver is not required to permanently monitor the road. There is no expectation of the user to respond to a request to intervene.
- **Full Driving Automation:** The system takes over speed and steering control completely and permanently, on all roads and in all situations. The driver sets a destination via a touchscreen. The driver cannot drive manually, because the vehicle does not have a steering wheel.

APPENDIX B (Continued)

Question	Unit/Coding
1. Have you read and understood the above instructions?	1 = Yes
2. The definitions given in the instructions are clear to me.	1= Disagree Strongly, 7= Agree Strongly
4. What is your primary mode of transportation?	1= Private Vehicle, 2= Public Transportation, 3= Motorcycle, 4= Walking, 5= Other
5. At what age did you obtain your first driver's license?	Year
6. On average, how often did you drive a vehicle in the last 12 months?	1=Never, 6=Every Day
7. About how many miles did you drive in the last 12 months?	1=0, 2=1-5000, …, 11= more than 50,000
8. Have you ever heard of the Google Driverless Car (Waymo) or other driverless cars?	2=No, 1=Yes
9. The idea of fully automated driving is fascinating.	1= Disagree Strongly, 7= Agree Strongly
10. Manual driving is enjoyable.	1= Disagree Strongly, 7= Agree Strongly
11. Partially Driving Automation will be enjoyable.	1= Disagree Strongly, 7= Agree Strongly
12. Conditional Driving Automation will be enjoyable.	1= Disagree Strongly, 7= Agree Strongly
13. Highly Driving Automation will be enjoyable.	1= Disagree Strongly, 7= Agree Strongly
14. Fully Driving Automation will be enjoyable.	1= Disagree Strongly, 7= Agree Strongly
15. Partially Driving Automation will be easier than manual driving.	1= Disagree Strongly, 7= Agree Strongly
16. Conditional Driving Automation will be easier than manual driving.	1= Disagree Strongly, 7= Agree Strongly
17. Highly Driving Automation will be easier than manual driving.	1= Disagree Strongly, 7= Agree Strongly
18. Fully Driving Automation will be easier than manual driving.	1= Disagree Strongly, 7= Agree Strongly
19. Some modern cars are equipped with Adaptive Cruise Control (a system that can automatically follow another car). How often did you use Adaptive Cruise Control (ACC) when driving in the last 12 months?	1=Never, 6=Every Day, (-1 = I do not have ACC, -2= I do not know what ACC is)
20. I would be comfortable driving in a fully automated driving vehicle without a steering wheel.	1= Disagree Strongly, 7= Agree Strongly
21. The idea of fully automated driving is silly. Scientists should focus on other, more important, research topics.	1= Disagree Strongly, 7= Agree Strongly
22. I believe that within 30 years, automated driving systems will be so advanced that it will be irresponsible to drive manually.	1= Disagree Strongly, 7= Agree Strongly

APPENDIX C

STRUCTURED INTERVIEW QUESTIONS

[read the framing instruction again]

- When you read this instruction at the beginning, what did you think about the vehicle?
- How did you think about the vehicle's performance?
- Do you think the car was driving in a safe manner?
- If you had the chance in real life, would you use the automated vehicle that would perform the same as the one you just saw in the study?
- Did you notice anything wrong with or strange about the vehicle during the study?

If Yes

-> What did you think while the error occurred?

-> What did you think might be the reason for the error?

->How did that affect your trust in the vehicle's capabilities?

->What did you think about the drives after the error?

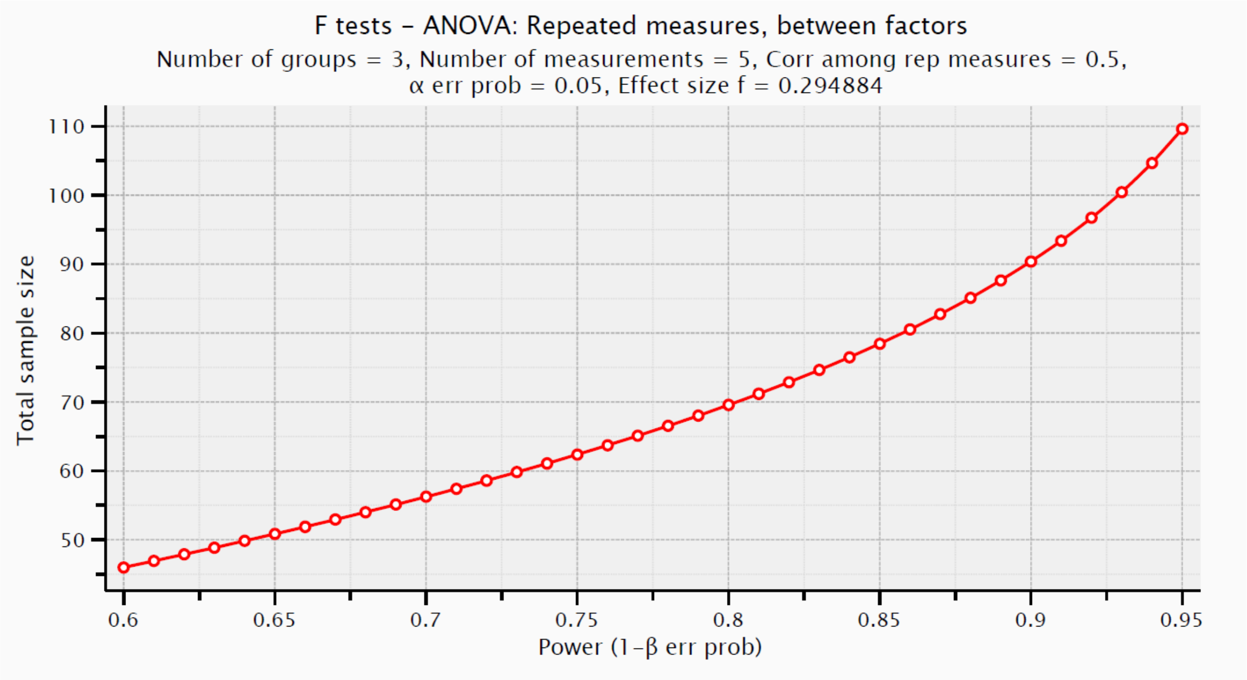
-> followup: did these drives make you think about the reliability of the vehicle?

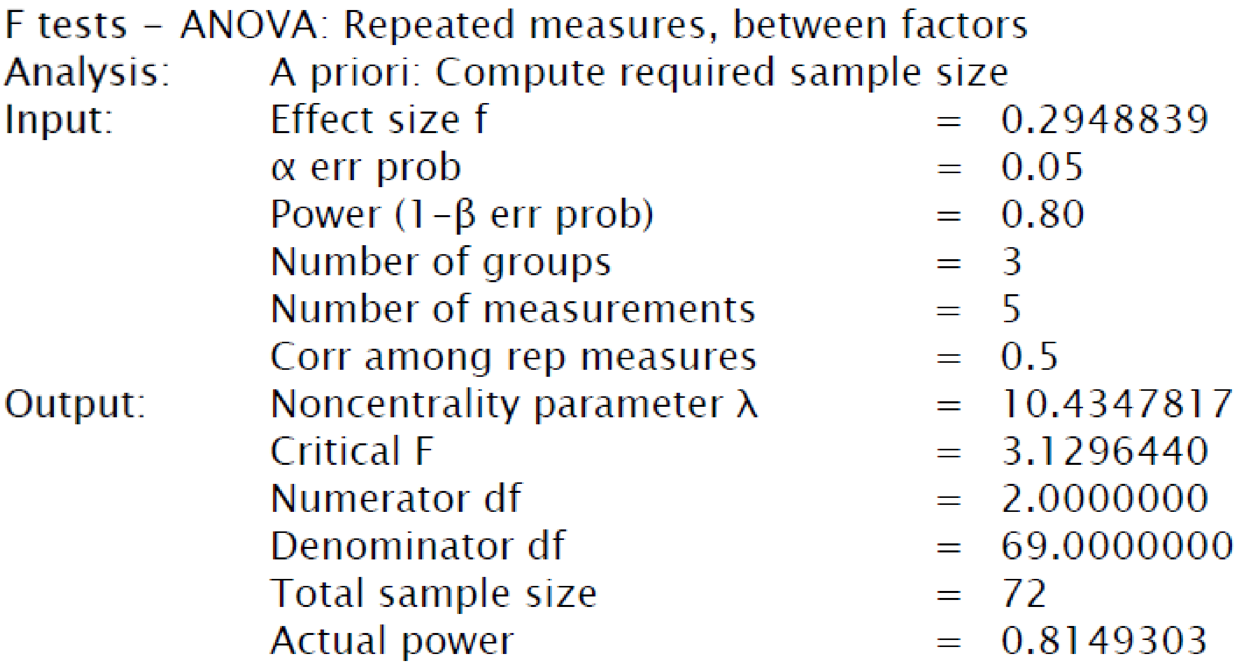
If No

- Show video of error
- Did you notice this during the drive?
    - If "Yes" ->
        - What did you think?
    - If "No" ->
        - What would you think if you had seen this error?
        - Why do you think you didn't notice this error?
- Do you consider this an error of the Driving Automation System?
- What did you think might be the reason for the error?
- How would this error affect your trust in the vehicle's capabilities?
- If you imagine drives after this error that do not have any errors, what would you think about those drives?
    - Followup: would these drives make you think about the reliability of the vehicle?

APPENDIX D

POWER ANALYSIS FOR EXPERIMENTS 2 AND 3





```
F tests – ANOVA: Repeated measures, between factors
Analysis:     A priori: Compute required sample size
Input:        Effect size f                        =   0.2948839
              α err prob                           =   0.05
              Power (1–β err prob)                 =   0.80
              Number of groups                     =   3
              Number of measurements               =   5
              Corr among rep measures              =   0.5
Output:       Noncentrality parameter λ            =   10.4347817
              Critical F                           =   3.1296440
              Numerator df                         =   2.0000000
              Denominator df                       =   69.0000000
              Total sample size                    =   72
              Actual power                         =   0.8149303
```

APPENDIX E

POST-STUDY 2 STRUCTURED INTERVIEW QUESITONS

[read the framing instruction again]

- When you read this instruction at the beginning, what did you think about the vehicle?
- How did you think about the vehicle's performance?
- Do you think the car was driving in a safe manner?
- If you had the chance in real life, would you use the automated vehicle that would perform the same as the one you just saw in the study?
- Did you notice anything wrong with or strange about the vehicle during the study?

-> What did you think while the error occurred?

-> What did you think might be the reason for the error?

->How did that affect your trust in the vehicle's capabilities?

->What did you think about the drives after the error?

-> followup: did these drives make you think about the reliability of the vehicle?

APPENDIX F

POST-STUDY 3 STRUCTURED INTERVIEW QUESITONS

[read the framing instruction again]

- When you read this instruction at the beginning, what did you think about the vehicle?
- Overall during the drive, what did you think about the vehicle's performance?
- Did you think the car was driving in a safe manner?
- Did you notice anything wrong with or strange about the vehicle during the study?

  -> Can you describe what you saw?

  -> Would you consider what you described an error?

  -> What did you think while the [error, event, strange situation, etc.] occurred?

  -> What did you think might be the reason for the [error]?

  -> How did that affect your trust in the vehicle's capabilities?

  -> [Remind of the post drive 3 message] – What did you think about the message from the automated driving system after the drive where the [error] occurred?

  -> How do you feel that this message influenced your trust in the system?

  -> What did you think about the drives after the error?

  -> followup: What did these drives make you think about the performance of the vehicle?

- If you had the chance in real life, would you use the automated vehicle that would perform the same as the one you just saw in the study?

VITA

Department of Psychology
Old Dominion University
Norfolk, VA 23529

Scott Anthony Mishler
smishler17@gmail.com
https://sites.google.com/view/scottmishler

EDUCATION

2019 – Expected 2023	**Ph.D., Psychology**, Old Dominion University, VA
2017 - 2019	**M.S., Psychology**, Old Dominion University, VA Thesis: *Whose drive is it anyway?: Using multiple sequential drives to establish patterns of learned trust, error cost, and trust repair while considering daytime and nighttime differences as a proxy for difficulty*
2016 - 2017	**Master's Student, Engineering Psychology**, New Mexico State University, NM First Year Project: *Driver's type of response to auditory car warnings in a semi-autonomous vehicle mediates safety*
2012 - 2016	**B.S., Psychology** - Purdue University, IN Focus: Human Factors, Cognitive Psychology Senior Research Project: *Can the working memory representation of a stimuli mediate the Simon effect through variation in color category*

SELECT PUBLICATIONS AND CONFERENCE PROCEEDINGS

Mishler, S., & Chen, J. (2023). Boring but Demanding: Using Secondary Tasks to Counter the Driver Vigilance Decrement for Partially Automated Driving. *Human Factors.*

Mishler, S., & Chen, J. (2023). Effect of automation failure type on trust development in driving automation systems. *Applied Ergonomics*, 106.

Chen, J., **Mishler, S.**, & Hu, B. (2021). Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Transactions on Human-Machine Systems*, *51*(5), 463-473.

Mishler, S., Jeffcoat, C., & Chen, J. (2019). Effects of Anthropomorphic Phishing Detection Aids, Transparency Information, and Feedback on User Trust, Performance, and Aid Retention. In *Proceedings of the Human Factors and Ergonomics Society 63rd International Annual Meeting*. Washington DC: HFES.

Chen, J., **Mishler, S.**, Hu, B., Li, N., Proctor, R. W. (2018). The Description-Experience Gap in the Effect of Warning Reliability on User Trust and Performance in a Phishing Detection Context. *International Journal of Human-Computer Studies,* 119, 35-47.

Mishler, S., Chen, J. (2018). Effect of response method on driver responses to auditory warnings in simulated semi-autonomous driving. In *Proceedings of the Human Factors and Ergonomics Society 62nd International Annual Meeting.* Washington DC: HFES.