Winter 2018

The Effects of Automation Transparency and Ethical Outcomes on User Trust and Blame Towards Fully Autonomous Vehicles

Nathan Andrew Hatfield
Old Dominion University

THE EFFECTS OF AUTOMATION TRANSPARENCY AND ETHICAL OUTCOMES ON USER TRUST AND BLAME TOWARDS FULLY AUTONOMOUS VEHICLES

by

Nathan Andrew Hatfield
B.S. May 2014, Christopher Newport University

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

MASTER OF SCIENCE

EXPERIMENTAL PSYCHOLOGY

OLD DOMINION UNIVERSITY
December 2018

Approved by:

Bryan E. Porter (Director)

YusukeYamani (Member)

Jeremiah Still (Member)

ABSTRACT

THE EFFECTS OF AUTOMATION TRANSPARENCY AND ETHICAL OUTCOMES ON USER TRUST AND BLAME TOWARDS FULLY AUTONOMOUS VEHICLES

Nathan Andrew Hatfield
Old Dominion University, 2018
Director: Dr. Bryan Porter

The current study examined the effect of automation transparency on user trust and blame during forced moral outcomes. Participants read through moral scenarios in which an autonomous vehicle did or did not convey information about its decision prior to making a utilitarian or non-utilitarian decision. Participants also provided moral acceptance ratings for autonomous vehicles and humans when making identical moral decisions.

It was expected that trust would be highest for utilitarian outcomes and blame would be highest for non-utilitarian outcomes. When the vehicle provided information about its decision, trust and blame were expected to increase. Results showed that moral outcome and transparency did not influence trust independently. Specifically, trust was highest for non-transparent non-utilitarian outcomes and lowest for non-transparent utilitarian outcomes. Blame was not found to be influenced by either transparency, moral outcome, or their combined effects. Interestingly, acceptance was determined to be higher for autonomous vehicles that made the same utilitarian decision as humans, though no differences were found for non-utilitarian outcomes.

This research draws on the importance of active and passive harm and suggests that the type of automation transparency conveyed to an operator may be inappropriate in the presence of actively harmful moral outcomes. Theoretical insights into how ethical decisions are evaluated when different agents (human or autonomous) are responsible for active or passive moral decisions are discussed.

This thesis is dedicated to my best friend, Maggie.

ACKNOWLEDGMENTS

There are many people who have contributed to the successful completion of this thesis. First, I would like to thank my advisor, Dr. Porter, for his many hours of guidance and support. His willingness to supervise a thesis that ventured into messy moral dilemmas is commendable; from start to finish, his contributions were invaluable.

Similarly, I would like to thank Drs. Yamani and Still for their feedback on the conceptual layout and structure of this manuscript. I would also like to thank Dr. Miller for his willingness to talk "trolleyology" with me, which ultimately gave me the peace of mind needed for the scenarios used in this thesis.

Lastly, I would like to pay special tribute to my family. Mom and Dad, thank you for trying to understand my gripes and helping me see the bigger picture. Maggie, thank you for your continual encouragement and unmerited love. You all have carried me a long way.

NOMENCLATURE

AV	Autonomous Vehicles
NHTSA	National Highway Traffic Safety Administration

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

I. INTRODUCTION

In recent news, fully autonomous vehicles have been faced with real-world moral scenarios similar to those posed by the trolley problem. As of the writing of this thesis, Uber has garnered national coverage after its self-driving vehicle failed to detect a pedestrian crossing the road, resulting in the death of the pedestrian. The self-driving vehicle was occupied with a passenger who experienced the vehicle's failure to swerve, in which it is plausible that the passenger's attitudes toward the self-driving vehicle—to include trust and blame— changed as a direct result of the incident. Similarly, a passenger traveling in Tesla's Model X engaged the adaptive cruise control feature during transport. While in an automated state, the vehicle conveyed multiple visual warnings to the driver before crashing into a barricade, causing the death of the single driver (Brown, 2018). Further inspection of the crash revealed that the driver did not reassume control of the vehicle prior to the collision. Both accidents indicate that autonomous vehicles have the potential to make costly errors with moral significance. Thus, understanding how conveyed system information in the context of moral outcomes can impact attitudes towards autonomous vehicles is critical.

The invention of the automobile has transformed the way humans travel and continues to be a growing area for applied psychological research (e.g., Adrian, Postal, Moessinger, Rascal, & Charles, 2011; Hennessy & Wiesenthal, 1999; Ross, et al., 2015; Stokols, Novaco, Stokols, & Campbell, 1978). The American Automobile Association Foundation for Traffic Safety (AAAFTS) reported that American drivers spend 46 minutes traveling each day, amounting to nearly 10,700 accrued miles annually (Triplett, Santos, & Rosenbloom, 2015). Since 1991, aggregate commute time and mileage have steadily increased (Travel Volume Trends, Federal

Highway Administration, 2016); as a result, social and economic risks—such as fatalities and carbon emissions—have increased (Blincoe et al., 2002; National Safety Council, 2017), garnering attention from policy makers, researchers, and others impacted by such risks.

The World Health Organization (2013) reported that 1.25 million fatalities in 2013 were a result of automobile accidents, making it the 9th leading contributor of death globally. In the United States alone, 2015 federal crash data revealed the largest year-to-year percent increase in motor vehicle deaths in nearly 50 years, resulting in 35,092 fatalities, and economic costs of $242 billion dollars (NHTSA, 2015). The National Highway Traffic Safety Administration (NHTSA)—after reviewing results from the National Motor Vehicle Crash Causation Survey—concluded that 93% of automobile accidents are a result of human error (Bellis & Page, 2008; see also Singh, 2015). With these factors in mind, automobile manufacturers and policy makers have pursued initiatives to offset the staggering number of life-threatening safety hazards, recently with the goal of minimizing the role of the human operator through the advent of autonomous vehicle technologies.

Autonomous Vehicle Technologies

Autonomous vehicle (AV) technologies have rapidly evolved in the 21st century, giving the driver more flexibility, while minimizing routine-driving tasks normally reserved for the human operator (Kiernan, 2015). AVs are expected to increase public safety by reducing automobile accidents related to deliberate and even unintentional hazardous driving behaviors (Fagnant & Kockelman, 2015; WHO, 2013). In addition, AV technologies have the potential to bolster cost savings up to $5,000 a year by reducing travel times, carbon emissions, and fuel consumption (Fagnant & Kockelman, 2015). These technologies are designed to navigate a

traffic environment with limited human intervention via a suite of onboard computing settings that analyze traffic-relevant data to make goal-oriented decisions on behalf of the driver.

To fully harness the benefits of AV technologies, NHTSA (2017) released a compendium of guidelines to educate state policy makers on how to safely integrate AVs for testing on public roadways. The guidelines feature a classification scheme for levels of automation adopted from the Society of Automotive Engineers (2016), ranging from: no automation (Level 0), driver-assistance automation (Level 1), partial automation (Level 2), conditional automation (Level 3), high automation (level 4) and full automation (Level 5).

After partitioning these levels in terms of human involvement, Levels 4 and 5 differ from Levels 0 through 3 by assuming all control of performance-based safety functions, with no expectation for the driver to intervene under any circumstance, though the option may be available. As a consequence, however, it is inevitable that fully autonomous vehicles (i.e., levels 4 and 5) will be exposed to crash scenarios whereby decisions must be made on behalf of the driver without the driver's knowledge or consent. Crash scenarios may have moral implications, like deciding who lives and dies in an unavoidable crash, and research is needed to determine how drivers will respond to such decisions carried out by fully autonomous vehicles.

Autonomous Ethics

Ethics is concerned with the governing principles or systems of thought that lead to subsequent actions and behaviors (Kiernan, 2015). Ethics, according to the Human Factors and Ergonomics Society (2005), is relevant to the safety of human operators, though research in this field is limited when it comes to fully autonomous vehicles (Kumfer, Levulis, Olson & Burgess, 2016). More recently, some attention has been given to the role of ethics in designing autonomous vehicles, with a specific emphasis on philosophical theories about right and wrong

actions (Kumfer, Levulis, Olson, & Burgess, 2016). For example, to ensure an occupant's safety, instances may arise in which autonomous vehicles will need to abandon the roadway to avoid accidents with other vehicles or pedestrians. Because fully autonomous vehicles are able to perform actions without human intervention, rapid moral decisions will need to be made independent of the driver. As expected, autonomous decisions pose real threats to drivers and innocent bystanders, while also introducing potential shifts in accountability for the consequences associated with the decision (Coeckelbergh, 2016).

According to Goodall's (2014) developmental work in autonomous vehicle ethics, scenarios that require ethical decisions from the autonomous vehicle have presented researchers with a series of challenges: first, to understand how to program vehicles to make ethical decisions that will not always have known outcomes; second, to ensure that cultural values and laws are embedded within the programming framework; and third, to confirm decisions are objective across a range of scenarios. These challenges will ultimately influence the way a vehicle decides to navigate potential accidents, as well as deciding who lives and dies in a fatal traffic scenario (Wallach & Allen, 2008). Greene, Rossi, Venable, and Tasioulas (2016) suggested that before embedding governing principles into autonomous agents it is critical to understand how to model such guidelines after human ethics, and thus referred to three schools of thought —deontology, utilitarianism, and virtue— to inform how autonomous systems ought to act. All three domains of thought have offered insight into machine ethics, though specific attention has been given to utilitarianism when considering the complex and often multi-layered scenarios that arise in traffic environments.

Utilitarianism, also known as consequentialism, is an approach to moral reasoning that places emphasis on maximizing an optimal decision; and moral decisions are right to the degree

in which the common good is promoted (Mill, 1901). Experimental ethics is a data-driven approach that enables researchers to quantify and understand moral intuitions. The Trolley Problem, originally designed by Foot (1978), has been used as such a method, and has helped researchers to better understand the cognitive and neural mechanisms involved in moral decision-making processes (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004), while also illuminating discussions about how to program autonomous vehicles (Nyholm & Smids, 2016).

The Trolley Problem has taken on many permutations since its inception, though the footbridge and switch scenarios have become more popularized in discussions about deontological (some actions are always wrong, regardless of their consequences) and utilitarian (an action is wrong in proportion to its consequence) ethics. The footbridge scenario describes a runaway trolley heading towards a group of people. Participants are presented with an option to save the group of people by pushing a man off of an overhanging footbridge and onto a track below, subsequently stopping the trolley and saving a group of five people in the trolley's onward path. The switch scenario is a modified version of the footbridge scenario and provides participants with the option of flipping a switch to divert the trolley onto a separate track on which one person is standing, ultimately taking the person's life. Both scenarios present participants with a trade-off between sacrificing one life to save a group of lives or allowing the group to perish by not intervening. Results have found that participants deem the switch option to be more morally acceptable than pushing a man off of a footbridge (Cushman, Young, & Hauser, 2006) due to a dichotomy that exists between personal and impersonal harm similar to Milgram's (1974) work. Though some have argued that the Trolley Problem is too far removed to make noteworthy contributions for understanding autonomous ethics (Nyholm & Smids,

2016), until more sophisticated methods become available (Cushman & Greene, 2012), the Trolley Problem remains a viable method to pioneer questions about driver attitudes towards autonomous vehicles.

Bonnefon, Shariff, and Rahwan (2016), for example, designed scenarios that mirrored the type of moral dilemmas featured in the Trolley Problem but revamped them to fit traffic scenarios including autonomous vehicles. Participants were surveyed to determine how ethical dilemmas influenced attitudes towards owning an AV. In general, participants thought it was morally acceptable for an AV to sacrifice a passenger if it meant pedestrians could be saved, though they were unwilling to purchase such an AV for themselves. These results highlight the value users place on autonomous systems to carry out decisions that align with users' ethical preferences, and also highlight how ethical decisions can differ based on personal or impersonal involvement with an autonomous vehicle (Greene et al., 2009).

Though Bonnefon et al. (2016) were the first to consider ethical dilemmas in terms of autonomous vehicle technologies from a survey standpoint, Wintersberger, Frison, Riener, and Boyle (2016) altered the switch Trolley Problem to better fit into a traffic context that was appropriate for a driving simulator. The switch scenario required participants either to select a switch in the driving simulator to reroute the vehicle thereby killing a group of pedestrians, or to refrain from selecting the switch whereby the vehicle would stay on its projected course, ultimately leading to self-inflicted harm. The participants were always presented with the option of selecting the switch to make the AV swerve, though information about the type and quantity of victim varied. In addition, each participant was given a percentage about their own life expectancy to help the researchers determine the maximum threshold value in which they would agree to stay the course and not swerve. The results revealed a significant effect of knowledge

about life expectancy on deciding to swerve or stay—with participants being willing to risk their own life to save the lives of others even when the probability of surviving the crash was low. In addition, as the number of potential pedestrian causalities increased, the likelihood of the participant allowing the vehicle to risk their life also increased.

Taken together, these studies reveal that data-driven techniques can be used to measure decisions about usually abstract thoughts, which can be used to explore attitudes towards AVs that make utilitarian or non-utilitarian decisions. The Trolley Problem may also provide a framework to explore the effects of design features, such as automation transparency, on user attitudes towards impending autonomous decisions. For example, if information about the vehicle's performance is withheld from the driver or if the vehicle's actions deviate from the driver's expectations, the driver may express different attitudes, such as trust and blame, towards the vehicle, potentially and significantly altering the driver-vehicle relationship.

Automation Transparency

An important factor of well-designed automation is the ability for automation to provide feedback to the user about its performance state. Disclosing information to the user about the system's purpose, underlying processes, intent, future actions, or reasoning processes is considered automation transparency (Endsley, Bolstad, & Jones, 2003; see also: Lee, 2012; Mercado et al., 2016, p. 402). Ultimately, automation transparency serves to furnish the user with an understanding of the automation's internal operations and logic (Seong & Bisantz, 2008), with the goal of supporting supervisory control or manual intervention (Ososky et al., 2014; Sanders et al., 2014). Automation transparency can eliminate speculation about poor performance originating from a system's design, leading to increased trust (Chen, Barnes, & Harper-Sciarini, 2011; Glass, McGuinness, & Wolverton, 2008; Lyons et al., 2017; Wang,

Jamieson, & Hollands, 2009), engagement (Wright, 2015; Lyons, 2013), and understanding (Körber, Prasch, & Bengler, 2018) when interacting with low and intermediate levels of automation (Ensdley, 2017).

Adaptive cruise control (ACC), for example, is an intermediate automation setting that enables the driver to select a desirable speed to be implemented by the vehicle's throttle-accelerator. ACC is not fully autonomous and therefore relies on the driver to shift between manual and intermediate automation settings. ACC collects data from the driving environment and makes automated speed adjustments to maintain appropriate distance from a lead vehicle. The system may request the driver to reassume control if performance limitations are exceeded. A driver's ability to understand an ACC system, to include knowing how and when to retake control, is critical for avoiding potential crashes. ACC operators generally have difficulty understanding how the system operates due to a lack of conveyed system information (Jenness et al., 2008; Körber, Prasch, & Bengler, 2018; Stanton & Marsden, 1996), potentially interfering with their ability to anticipate and retake control of the vehicle in a skilled manner (see Seppelt & Lee, 2007). In contrast, ACC systems that display a continuous stream of real-time information about the system's performance can assist the operator with calibrating appropriate user reliance, intervention and following strategies, and faster brake response times compared to ACC systems that do not provide such information (Seppelt & Lee, 2007).

Trust in Automation

Trust in automation has been investigated in many contexts and has been found to mirror social trust (Nass, Fogg, & Moon, 1996). Trust in automation is best captured during scenarios characterized by uncertainty and vulnerability (Lee & See, 2004), whereby the user evaluates the automation based on predefined expectations of how the automation is expected to act. Several

models of trust have been proposed (Barber, 1983; Lee & Moray, 1994; Muir, 1994), with Ajzen and Fishbein's (1980) model being careful not to conflate distinct constructs of trust, such as beliefs, intentions, attitudes, and behaviors. In fact, it is on the basis of these theoretical components that Lee and See (2004, p. 54) defined trust as "an attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Lee and See (2004) re-emphasized Lee and Moray's (1992) three-dimensional model of trust that includes: performance, process, and purpose. The performance axiom of trust rests on a system's ability to execute actions in a manner that is consistent with a user's goals; the process axiom of trust rests on a system's capability for completing an action given certain situational contexts; and the purpose axiom of trust rests on a system's intentionality for why it was developed.

Research on trust in automation has also considered the interplay of a system's reliability and human's expectation of how a system will act (see Hoff & Bashir, 2015; Muir & Moray, 1996). Moreover, human-operators tend to trust automated systems that are highly reliable (Lee & Moray, 1992) and behave according to expectations (Merritt & Ilgen, 2008; Robert, Denis, & Hung, 2009), and distrust automated systems that fail to meet expectations (Lee & See, 2004). Though many frameworks of trust in automation exist, this thesis operated on Lee and See's (2004) definition of trust and evaluated user's attitudes towards an autonomous vehicle.

The level of trust humans place in automation can also influence strategies on how to use automation (Lee & Moray, 1994). When humans overtrust a system to perform decisional tasks more than themselves, they can become complacent and disengaged, leading to automation bias (Mosier et al., 1998; Skitka, Mosier, & Burdick, 1999), poor detection of automation failures (Parasuraman, Molloy, & Singh, 1993), and automation misuse (Parasuraman & Riley, 1997). Similarly, when automation fails to live up to performance expectations, undertrust and

automation disuse can occur (Parasuraman & Riley, 1997), resulting in unnecessary manual control (Lee & Moray, 1994). Calibrating an appropriate level of trust is therefore essential for furnishing cooperative and productive relationships between humans and automation (Sheridan & Henessy, 1984), and is necessary to curtail misuse and disuse of fully autonomous vehicles (Lee & See, 2004). When the user is made aware of the system's inner workings, research suggests that trust is better calibrated, and can withstand and recover from automation failure (Wang, Jamieson, & Hollands, 2011).

Relationship Between Transparency and Trust

More recently, research has focused on the relationship between automation transparency and trust, but not explicitly in the context of fully autonomous vehicles (Levels 4 and 5). Verberne, Ham, and Midden (2012), for example, examined whether trust and acceptance ratings in ACC systems (Levels 2 and 3) varied as a function of the system's level of transparency. Research has revealed that transparency can be broadly operationalized as the presence or absence of information (e.g. Lyons et al. 2017). For example, Sheridan and Verplank's (1978) 10-level automation classification scheme can be understood in terms of transparency, with low automation (levels 1-5) providing information to the user without taking action on the user's behalf, intermediate automation (levels 6-7) providing information to the user while also taking action on the user's behalf, and full automation (levels 8-10) taking over complete control without providing information to the user about its actions. Verberne and colleagues (2012) found that ACC systems that provide information about driving tasks in the form of pictorial icons (information present) are more trusted and accepted than ACC systems that do not provide information (information absent). Moreover, trustworthiness was higher for ACC systems that

provided information to the user without taking action, an important finding when considering fully autonomous vehicles.

Similarly, Beller, Heesen, and Vollrath (2013) investigated whether changes in the driver-vehicle relationship occurred when uncertainty warnings about the ACC's performance were issued to the driver via a graphical modality displayed on the dashboard. The icon was used to give the driver an indication of the vehicle's uncertainty in times of inadequate self-performance, a form of automation transparency (Mercado et al., 2016). Uncertainty warnings resulted in higher subjective trust scores compared to non-warnings. Other research suggests that using textual (Chen et al., 2014) or auditory (Forster, Naujoks, & Neukum, 2017) explanations can increase trust so long as the system's decision-making processes are conveyed to the user (see Northdurft & Minker, 2016; Oduor &Wiebe, 2008; Sanders, Wixon, Schafer, Chen, & Hancock, 2014).

Blaming Automation

It is plausible that shared social norms between technology and humans (Nass, Fogg, & Moon, 1996) will lead to instances where the user will not only trust (Nass & Lee, 2001; Reeves & Nass, 1996) a system but also blame a system for not conforming to performance expectations, especially if nonconformities are perceived to be intentional (Malle, Monroe, & Guglielmo, 2014). Blame evaluations, according to Malle, Guglielmo, and Monroe (2012), can be divided into three categories: affirming norms, evaluating events, and evaluating agents. Affirming norms is the process by which norms are evaluated against other norms; evaluating events is the process by which observers appraise the outcome of an event against a set of norms; and evaluating agents is the process by which observers assign moral consequences, such as

blame, to agents that contributed to a negative event outcome (Malle, Guglielmo, & Monroe, 2014).

Before blame can be ascribed to the agent for a negative outcome, the individual first considers whether the outcome was intentional (Malle, Guglielmo, & Monroe, 2012). If the negative outcome is deemed to be intentional, the individual then determines whether the agent was justified in making the decision based on the reasons leading up to the outcome—whereby blame is less severe if the agent's reasons are justifiable (Malle, Guglielmo, & Monroe, 2014). Intentionality is a primary criterion for which blame is assigned to an agent (Alicke, 2000), and research suggests that humans consider agents, such as computers, to be intentional (Friedman & Millet, 1995).

In terms of automation research, blame has received little attention, but may be intimately related to trust in automation. Research has found that automation transparency, when presented in the form of system reasoning, increases subjective trust ratings (Glass, McGuiness, & Wolverton, 2008). Because automation transparency conveys a system's reasoning to the operator, the operator may also be in a position to evaluate the system's reasoning against the system's performance outcomes and assign blame when optimal outcomes are violated. In fact, Gray and colleagues (2007) found that technology with a conveyed state of mindfulness can make users more prone to blame systems for wrong actions, giving credence to the system being perceived as a responsible moral agent (Waytz, Heafner, & Epley, 2010). The relationship between automation transparency, trust, and blame therefore introduces the possibility of an automation irony (see Bainbridge, 1983). Specifically, automation transparency can be used to communicate a system's reasoning processes: leading to increased trust. In contrast, automated systems without transparency can make users less trusting, but the system should not be

perceived as an intentional agent with an independent mind worthy of blame to the same degree as a transparent system. Therefore, designing systems that are more transparent may increase trust while simultaneously increasing the opportunity for blame due to increased intentionality.

Relationship Between Transparency and Blame

Automated systems that possess anthropomorphic qualities, such as rational thought and expressive emotion, have been found to affect the way users trust and blame technologies (Gray, Gray, & Wegner, 2007). Waytz, Heafner, and Epley (2010), for example, investigated user-trust and blame assignment for anthropomorphic and non-anthropomorphic autonomous vehicles. The results found that anthropomorphic vehicles that possessed a name, gender, and voice were better trusted than vehicles that did not have such qualities. The study also found that participants were less likely to assign blame to anthropomorphic autonomous vehicles that were struck by another vehicle than agentic vehicles. Though the authors were unable to pinpoint why blame scores differed between autonomous vehicle conditions, they drew on the role human-like qualities played in diminishing the perceived responsibility attributed to the anthropomorphic vehicle. If the anthropomorphic autonomous vehicles were at fault for the accident, however, anthropomorphic qualities might also elicit more blame from the user due to perceived reasoning abilities, a precursor for ascribing blame (Malle, Guglielmo, & Monroe, 2014).

Similarly, Kim and Hinds (2006) explored blame and credit assignment in low and high autonomous robots. The study hypothesized that highly autonomous robots would receive higher blame scores than low autonomous robots, due to the perceived level of intention inherent in highly autonomous robots, such as conveyed information about performance and actions. Robots were considered transparent if an explanation was given for an unexpected behavior and non-transparent if not. The results confirmed that higher blame scores were assigned to highly

autonomous robots as compared to low autonomous robots. The results did not find the transparent manipulation to explain blame beyond the automation condition, arguably because the autonomy manipulation was already inherently transparent or non-transparent. These results indicate that fully autonomous robots might be perceived as transparent and therefore worthy of blame when performance information is conveyed to the user.

Malle, Schuetz, Arnold, Voiklis, and Cusimano (2015) investigated the degree to which blame assignment differed between a human or robot character when presented with the same moral dilemma. The dilemma was a classic utilitarian tradeoff in which a human or robot took action or remained inactive when determining the direction of a trolley. Acceptance ratings for taking action to produce a utilitarian outcome were higher for the robot than the human, suggesting that participants expected robots to make these decisions slightly more than humans. Conversely, when the robot character did not intervene to produce a utilitarian outcome, participants assigned more blame to the robot than the human character that made the exact same decision. The results highlight differences in moral expectations for humans and robots, with robots being held more accountable than humans for not intervening to produce a utilitarian outcome and also being held less accountable than humans for intervening to produce a utilitarian outcome. Moreover, though not measured in this study, the authors suggested that trust might be affected when robots deviate from expected outcomes, in which robots will need to offer reasoning for their decisions to maintain trust.

Human Factors, Ethics, and Autonomous Vehicles

Autonomous vehicles that are forced to make ethical decisions might be a strong avenue for exploring whether blame and trust assignment varies as a function of high or low transparency. Because humans and automated technologies share social norms during interaction

(Reeves & Nass, 1996), understanding the types of norms humans expect autonomous vehicles to uphold warrants the need for further investigation and presents a unique avenue for marrying ethics and human factors research.

Where prior research has looked at trust and blame separately, this thesis combined both constructs and included morality and transparency in the experimental design. Moreover, this thesis served to determine how autonomous vehicles are evaluated when making optimal or non-optimal moral decisions. This was an important extension to past research, in that evaluations would not only be made on the basis of a moral outcome but also the agent performing the outcome. Participants who are more morally accepting of optimal moral decisions were expected to be more trusting of autonomous vehicles that executed optimal moral decisions, especially when the autonomous vehicle's rationale supported such decisions. Moreover, rationale for optimal moral decisions might further increase trust than optimal moral decisions that lack rationale. Including a layer of system transparency in the experimental methodology may further support what is known about trust in automation. Yet, it remains unknown how system transparency will also influence blame ascription when non-optimal moral outcomes occur. This thesis examined potential changes in blame ascription when an autonomous vehicle was directly responsible for a moral outcome, dissimilar from Waytz, Hefner, and Epley's (2014) methodology that assessed blame when another vehicle was at fault.

Lastly, this thesis sought to explore potential discrepancies between Malle et al's (2015) and Bonnefon et al.'s (2015) findings. Bonnefon et al's (2015) research indicated that moral permissibility is highest for utilitarian autonomous vehicles that do not sacrifice the driver. Malle et al's (2015) research found that moral preference differs for humans and robots faced with the same moral scenario, with robots being expected to make utilitarian decisions more than humans.

Malle's experiment separated robot actions from human actions, whereas the nature of Bonnefon's study requires autonomous vehicles to perform actions on behalf of humans. Thus, exploring potential differences in moral permissibility when a human does or does not have control of the vehicle was expected to introduce dissimilarities in moral norms.

Therefore, the current study examined the following hypotheses:

Hypothesis 1

Trust and blame scores would differ between the transparent condition and the non-transparent condition, with trust and blame scores being higher in the transparent condition.

Hypothesis 2

Trust and blame scores would differ between utilitarian outcomes and non-utilitarian outcomes, with trust scores being higher in the utilitarian condition and blame scores being higher in the non-utilitarian condition.

Exploratory Interaction

Trust scores would decrease in the non-utilitarian non-transparent condition and increase in the utilitarian transparent condition. In addition, blame scores would increase in the non-utilitarian transparent condition and decrease in the utilitarian transparent condition.

Exploratory Question:

Moral acceptance ratings would be higher, in general, for AV outcomes than for synonymous human outcomes.

II. METHOD

Sample Estimate and Experimental Design

An *a priori* power analysis was conducted using a statistical power analysis software, G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007). The analysis assumed a power of .80 for the main effects and interaction. Given the 2 X 2 Analysis of Variance being conducted for trust and blame, results from the power analysis for the main effects and the interaction term were the same, due to the same degrees of freedom for each term. A conservative effect size of .05 partial eta squared was chosen for this study, with an alpha level set at .05. The power analysis indicated 152 participants were needed to detect the proposed effect. Thus, a minimum of 38 participants were needed for each condition of the 2 (transparent vs non-transparent) X 2 (utilitarian vs non-utilitarian) between-subjects design.

Participants

Undergraduate psychology students (N = 186, Mage = 23.5 years, SD = 6.57, age range: 18-51 years) were recruited from Old Dominion University's SONA participant pool: male (N = 43, Mage = 22.54) and female (N = 142, Mage = 23.20). Each participant was awarded .5 research credit after completing the study (for work expected to require less than 30 minutes). The only criterion for participation in this study was for students to have a valid driver's license and be at least 18 years of age.

Materials

Two scenarios provided the framework for each of the four separate conditions featured in this thesis (see Appendix A for each scenario). Following Greene, Morelli, Lowenberg, Nystrom, and Cohen's (2008) and Malle et al.'s (2015) methodology, we required participants to

read about an autonomous vehicle faced with a moral dilemma in which the autonomous vehicle made a utilitarian or non-utilitarian decision. Participants were then instructed to answer a moral acceptance scale based on the autonomous vehicle making a utilitarian decision. The moral acceptance scale required participants to answer "yes" or "no" about the moral acceptability of the autonomous vehicle, and were then required to rate the moral acceptability of the action on a nine-point Likert scale (1 = *completely unacceptable*, 9 = *completely acceptable*). At the end of the study, participants rated the moral acceptability of a similar moral dilemma, only this time a human driver was making the moral decision and not an autonomous vehicle.

Trust in automation was measured by a 12-item questionnaire developed by Jian, Bisantz, and Drury (2000) (see Appendix B). Each item was measured using a seven-point Likert scale (1 = *totally disagree*, 7 = *totally agree*). Each item of the scale was adapted so that the word "system" was replaced with "autonomous vehicle", similar to that of Verberne, Ham, Midden's (2012) questionnaire. Responses for each item (e.g., "I would be confident in the autonomous vehicle if it were my own.") were recorded so that higher scores represented higher trust in automation. According to the reliability analysis conducted, the data revealed moderate reliability (Cronbach's alpha = .77).

Blame was assessed using Waytz, Heafner, and Epley's (2014) *Blame for vehicle measure.* The measure is comprised of 8 items using a ten-point Likert scale (0 = *not at all*, 10 = *very much*) (see Appendix C). Questions measured participants' blame for the transparent or non-transparent vehicle's decision as well as the passenger of the vehicle (e.g., "How much is the car itself at fault?"). In total, six of eight items constituted a single composite score for blame ratings toward the self-driving vehicle, indicating strong internal consistency reliability (Cronbach's alpha = .87).

Procedure

Before soliciting participants, the study was approved by Old Dominion University's Institutional Review Board (IRB). Participants registered to complete the experiment via the department's SONA platform; the survey itself was administered via Qualtrics. Participants were given a brief overview of the study's purpose and a notification sheet describing the study. The experiment lasted approximately 15 minutes. Qualtrics randomly assigned participants to the conditions of a 2 (transparency level: transparent vs non-transparent) x 2 (driving outcome: utilitarian vs non-utilitarian) between-subjects experimental design, and counterbalanced the presentation of one of four driving scenarios. Each driving scenario included a brief description of the autonomous vehicle's capabilities before further manipulations (i.e., transparency or moral outcome) were made. Participants first provided a moral acceptance rating for the autonomous vehicle's initial moral dilemma, and then continued reading the scenario in which transparency and moral outcome were manipulated. For each moral outcome, participants were presented with a manipulation check in which they were asked if the autonomous vehicle provided any information about why it decided to pursue a particular course of moral action. After reading the final outcome, participants' attitudes of blame and trust were measured. Participants were then asked to complete the same moral acceptance scale but for a human who was faced with the exact same dilemma as the autonomous vehicle. Upon completing all three scales, demographic information was collected from each participant (see Appendix D).

III. RESULTS

Of the 186 participants, 35% incorrectly answered the manipulation check. The manipulation check was designed to assess participants' attention to each scenario and understanding of the vehicle's conveyed information. For the transparency condition, the vehicle provided information to participants about its course of action, in which participants should have indicated that the vehicle did in fact provide information. Similarly, for the non-transparent condition, participants should have answered "no" when asked if the vehicle provided information about its course of action, because no information was provided to the participant. Removing participants who failed the manipulation check from the sample resulted in unequal sample sizes across conditions. Therefore, the results and conclusions were interpreted with caution for those who incorrectly answered the manipulation check.

All analyses were conducted with a standard alpha level set at .05, using R (R Core Team, 2017). Data for blame and trust were analyzed separately in a 2 (transparency, no transparency) x 2 (utilitarian, non-utilitarian) between-subjects ANOVA. The data were first analyzed to ensure the assumptions of ANOVA were not violated. Subsequently, all effects—main and interaction—were tested, and Bonferroni's adjustment was used for post-hoc analyses to maintain familywise alpha at .05.

Table 1

Means, standard deviations, and correlations with confidence intervals

Variable	*M*	*SD*	1	2	3
1. Mean blame	6.97	2.47			
2. Mean trust	3.29	0.91	.27** [.13, .39]		
3. Human accept	3.15	2.22	-.18* [-.31, -.04]	.11 [-.04, .25]	
4. AV accept	3.34	2.20	-.19* [-.32, -.04]	.01 [-.14, .15]	.71** [.63, .77]

Note. * indicates $p < .05$; ** indicates $p < .01$. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation.

Trust

Average mean trust scores were submitted to a 2 x 2 between subjects ANOVA, with transparency and moral outcome as fixed, between subjects factors. Mean trust scores were normally distributed as evidenced by skewness (.220) and kurtosis (.785) values falling below recommended cutoff values of $\pm$ 2 (Gravetter & Wallnau, 2014). Levene's test indicated that homogeneity of variance was not violated, $F(3, 182) = 1.84$, $p = .142$. Averaging across moral outcome, a significant main effect for transparency was not found, $F(1, 182) = .657$, $p = .419$, partial $\eta^2 = .004$, indicating that a vehicle communicating its rationale may not increase trust any more than a vehicle not communicating its rationale, refuting *Hypothesis 1* that argued more communication would increase trust. Averaging across transparency, there was no significant main effect for moral outcome, $F(1, 182) = 2.29$, $p = .132$, partial $\eta^2 = .012$, suggesting that

utilitarian moral decisions may not increase trust any more than non-utilitarian moral decisions, refuting *Hypothesis 2* that argued utilitarian decisions would increase trust more than non-utilitarian decisions. The interaction of transparency and moral outcome was significant $F(1, 182) = 4.96$, $p = .027$, partial $\eta^2 = .03$. Post-hoc comparisons using Bonferroni's familywise adjustment (see Maxwell & Delaney, 2003, p. 308) explored the interaction by examining the simple effect of A within specific levels of B ($\alpha = \frac{.05}{2} = .025$). Post hoc analyses revealed that mean trust scores for the non-utilitarian non-transparent condition ($M = 3.59$, $SD = 1.06$) were significantly higher than the utilitarian non-transparent condition ($M = 3.10$, $SD = .86$), $t(90) = -2.44$, $p = .017$, $d = .51$. Such findings indicate that differences in trust for non-utilitarian moral decisions depends on the absence of communication to the driver. This finding departed from the proposed direction of the *Exploratory Interaction* that transparent AVs would receive more trust when making utilitarian decisions than non-transparent AVs making non-utilitarian decisions. No other significant simple effects were observed.

Table 2

Fixed-Effects ANOVA results using Mean Trust as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 95% CI [LL, UL]
(Intercept)	2019.10	1	2019.10	2511.00	<.001		
moral	1.84	1	1.84	2.29	.132	.01	[.00, .06]
transp	0.53	1	0.53	0.66	.419	.00	[.00, .04]
moral x transp	3.98	1	3.98	4.96	.027	.03	[.00, .09]
Error	146.35	182	0.80				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

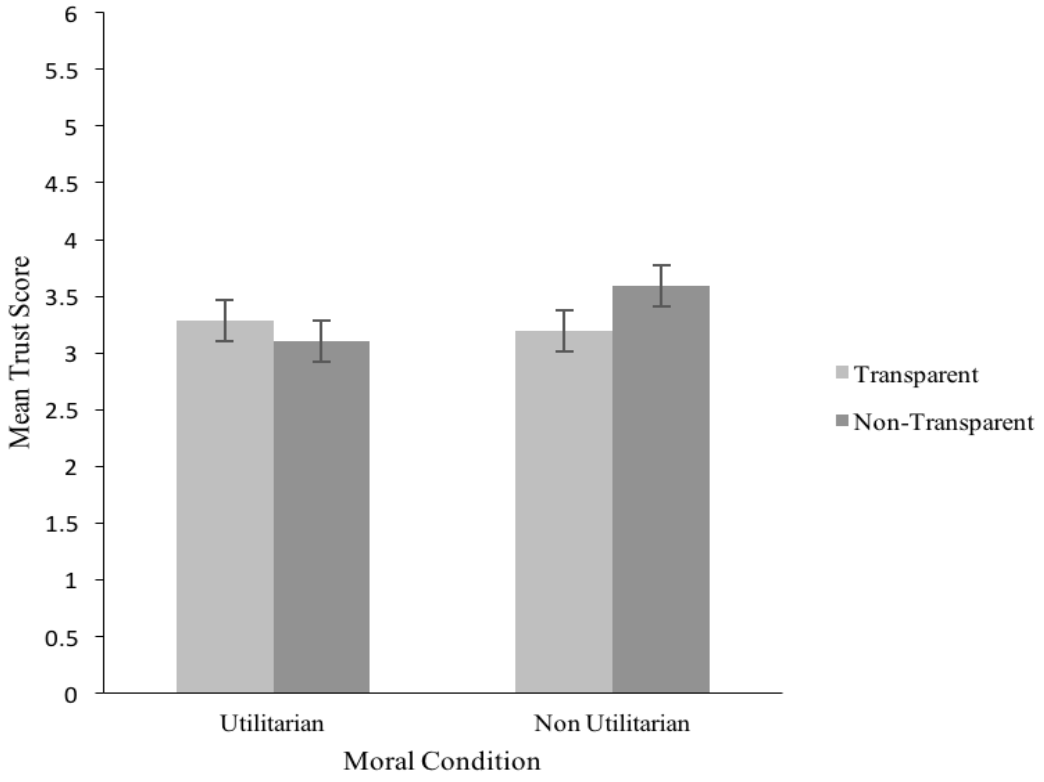


Figure 1. The difference in mean trust scores between utilitarian and non-utilitarian conditions as a function of transparency. Error bars represent 95% between-subject confidence intervals.

Table 3

Bonferroni Comparison for Mean Trust

Comparisons	Mean Trust Difference	Std. Error Difference	*p*	95% CI of the Difference	
				Lower Bound	Upper Bound
Trans vs Non-Trans (level: Util)	.19	0.18	.311	-0.17	0.55
Trans vs Non-Trans (level: Non-Util)	-.39	0.19	.038	-0.78	-0.02
Util vs. Non-Util (level: Trans)	.09	0.17	.582	-0.24	.43
Util vs. Non-Util (level: Non-Trans)	-.49	0.20	.017*	-0.89	-.09

* $p < 0.025$ (Bonferroni correction for significant interaction requires an examination of the simple effect of A within specific levels of B and vice versa. To maintain familywise alpha at .05: $\alpha = \frac{.05}{b} = \frac{.05}{2} = .025.$).

Blame

Average mean blame scores were submitted to a 2 x 2 between subjects ANOVA, with transparency and moral outcome as fixed, between-subjects factors. Mean blame scores were normally distributed as evidenced by skewness (-.936) and kurtosis (.343) values falling below recommended cutoff values of ± 2 (Gravetter & Wallnau, 2014). Levene's test indicated that homogeneity of variance was not violated, $F(3, 182) = .182$, $p = .909$. All of the proposed effects were not significant, refuting *Hypotheses 1*, *Hypothesis 2*, and the *Exploratory Interaction*. These findings suggest that utilitarian and non-utilitarian moral outcomes—as well as the vehicle's communication strategy—do not influence blame towards autonomous vehicles independently or jointly.

Table 4

Fixed-Effects ANOVA results using Mean Blame as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{\text{partial}}\eta^2$	$_{\text{partial}}\eta^2$ 95% CI [LL, UL]
(Intercept)	9022.27	1	9022.27	1464.02	<.001		
moral	5.99	1	5.99	0.97	.326	.01	[.00, .05]
transp	0.30	1	0.30	0.05	.825	.00	[.00, .02]
moral x transp	0.17	1	0.17	0.03	.867	.00	[.00, .02]
Error	1121.60	182	6.16				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

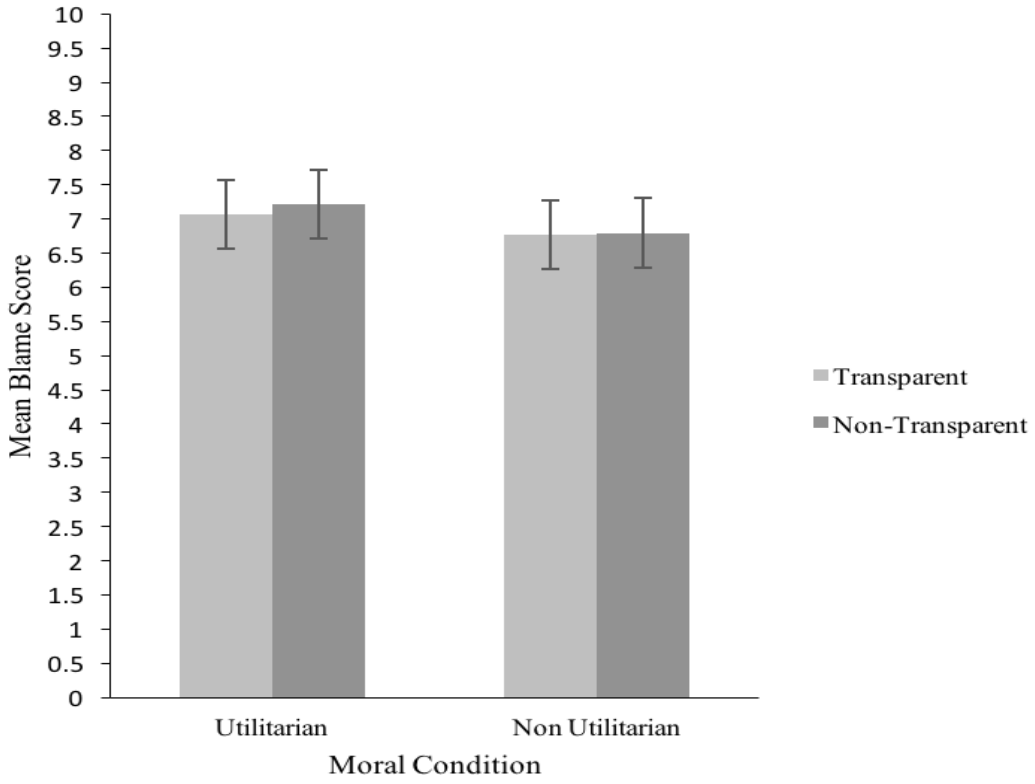


Figure 2. The difference in mean blame scores between utilitarian and non-utilitarian conditions as a function of transparency. Error bars represent 95% between-subject confidence intervals.

Acceptance

Lastly, two paired-samples t-tests were conducted to compare moral acceptance scores towards human drivers and autonomous vehicles for each moral outcome type. The results indicated that for utilitarian outcomes, mean moral acceptance scores were significantly higher for autonomous vehicles ($M = 3.33$, $SD = 2.2$) than for humans ($M = 2.9$, $SD = 2.16$), $t(93) = 2.75$, $p = .007$, $d = .28$. This finding suggests that humans may hold other humans to higher moral standards than vehicles when an actionable decision must be made. Moral acceptance scores for autonomous vehicles ($M = 3.35$, $SD = 2.21$) were not significantly different from humans ($M = 3.39$, $SD = 2.27$) when making a synonymous non-utilitarian decision, $t(91) = -.23$, $p = .82$, $d = .02$, indicating that moral standards may be similar for both agents when the

outcome is a result of inaction. See figures 3 and 4, respectively. Taken together, the results partially support the *Exploratory Question* that AVs would receive higher acceptance ratings than humans for synonymous moral outcomes.

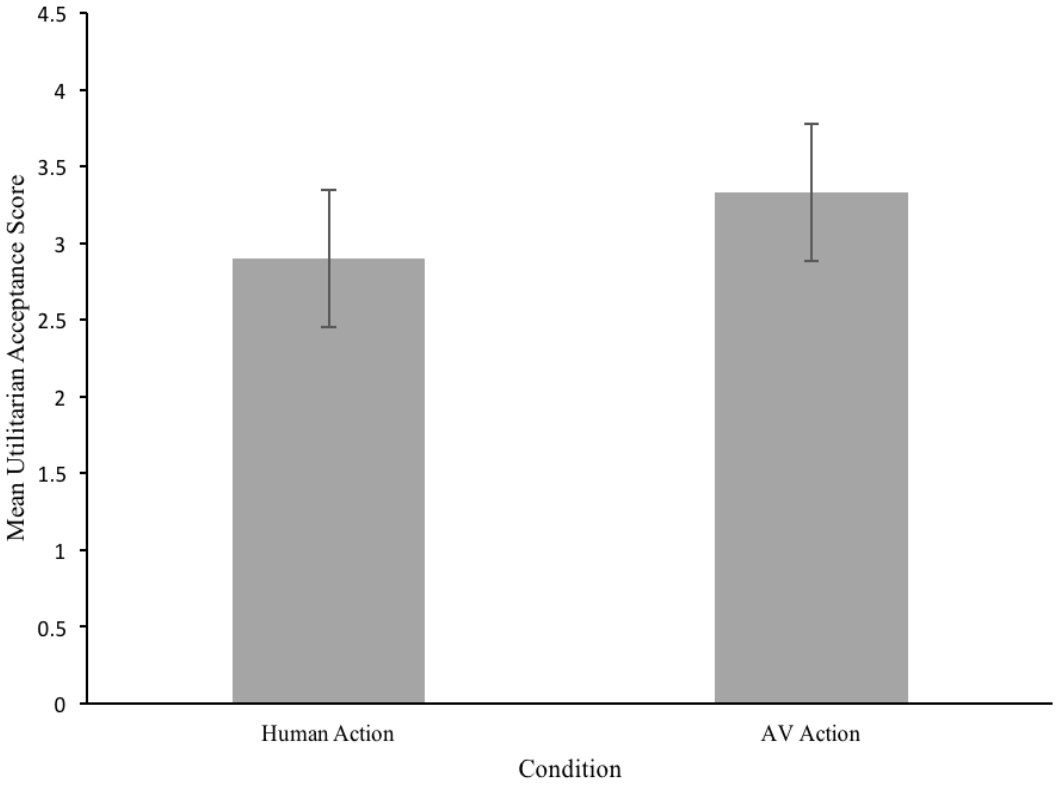


Figure 3: *Note.* The difference in mean acceptance scores between humans and autonomous vehicles making the same utilitarian decision. Error bars represent 95% confidence intervals.

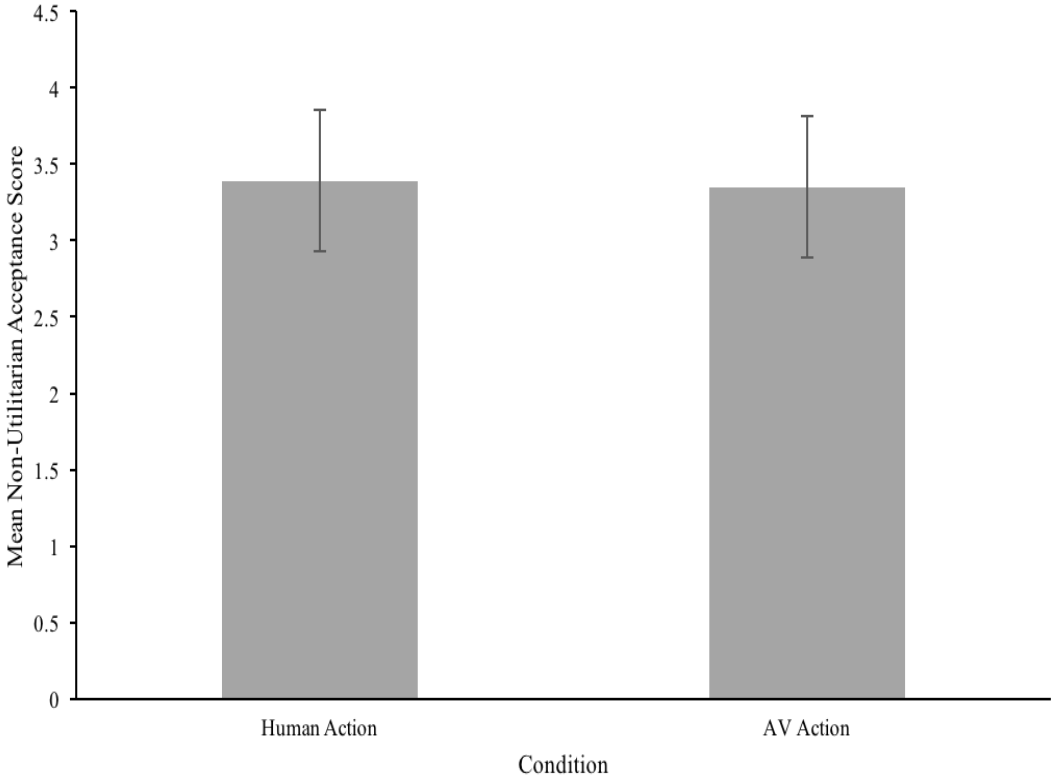


Figure 4: *Note.* The difference in mean acceptance scores between humans and autonomous vehicles making the same non-utilitarian decision. Error bars represent 95% confidence intervals.

IV. DISCUSSION

The effect of autonomous vehicle transparency on trust and blame has not been investigated in moral contexts, where conveyed information may or may not interact with a moral outcome. To explore these constructs, the current study required participants to read through a moral scenario that resulted in an autonomous vehicle making a utilitarian or non-utilitarian decision with or without information to support its decision.

Overview of Findings and Theoretical Implications

Trust in automation has been found to vary as a function of transparency (Lyons et al., 2017; Verberne, Ham, & Midden, 2012), with trust increasing for systems that convey information to the user during performance. Blame has also been found to increase in the presence of automation transparency (Kim & Hinds, 2006). Unexpectedly, automation transparency did not result in significantly higher trust or blame scores independent of moral outcome, contesting Hypothesis 1. A potential lack of manipulation saliency may have interfered with the expected transparency effect when acting independently. When, however, the transparency manipulation was coupled with non-utilitarian outcomes, trust in automation increased (Exploratory Interaction). Subsequently, the effects of transparency and moral outcome did not influence blame as expected. It is difficult to pinpoint an explanation for these results—because blame ascription, in part, requires an agent to intentionally violate a user's expectation, and moral expectations were not explicitly balanced in the experimental procedures—a limitation of the study.

Second, utilitarian moral outcomes were expected to result in higher user trust and lower user blame, due to partiality given to five-for-one tradeoffs that mediate moral action through some other device (in this case the vehicle) (Hauser, Cushman, Young, Jin, & Mikhail, 2007;

Moore, Clark, & Kane, 2008). The results do not support this hypothesis and indicate that moral outcomes may not be assessed solely on aggregate loss and gains, but rather other factors, such as intentionality (Greene et al, 2009) and direct or indirect harm (Aquinas, 1988; Greene et al., 2001).

Literature on the Doctrine of Doing and Allowing (Quinn, 1989; Rickless, 1997) may provide insight into why higher trust was found for non-utilitarian outcomes in the presence of non-transparency. The automaticity hypothesis suggests that moral actions are evaluated implicitly, in which active harm is less morally acceptable than passive harm (Cushman et al., 2006; Gleicher et al., 1990; Spranca, Minsk, & Baron, 1991). The practice of allowing harm to occur is known as an omission bias, whereas doing harm is known as a commission bias (Baron & Ritov, 2004). Engaging in an act that results in a negative outcome as a means to prevent a more widespread negative outcome is more morally salient, and thereby less morally acceptable (Ritov & Baron, 1990). Moreover, a vehicle remaining on course and allowing five pedestrians to die is more acceptable (omission bias) than choosing to kill one pedestrian actively (commission bias), because adverse emotions are more salient for action than for inaction (Bartels, 2008; Greene et al., 2008).

Though the optimal decision in a moral dilemma would be to ensure the least amount of aggregate loss, the vehicle would have to actively inflict harm in order to do so, perhaps activating strong emotional responses and perceived responsibility for such decisions (see Greene et al., 2001; Ritov & Baron, 1990). Considering that both outcomes resulted in death, it is plausible that participants supported the vehicle's omission, leading to higher trust than vehicles that made a commission (active response), thereby supporting the Doctrine of Doing and Allowing and discrediting traditional consequentialist beliefs. Taken together, consequentialism

should not serve as a panacea for autonomous vehicle programming guidelines but rather a starting point. To increase trust, manufacturers should better understand the diverse nature of human moral preferences and how autonomous systems can be designed to execute actions in alignment with such preferences.

The transparency manipulation in this thesis can be understood in terms of messages that contain *how* information or a combination of *how* and *why* information. Participants assigned to the transparency condition, for example, were provided with information about *how* and *why* the autonomous vehicle took a specific course of action (*how and why* manipulation), while participants in the non-transparent condition were only informed about the vehicle's course of action without the vehicle's rationale (*how* without *why* manipulation). Paralleling Koo and colleague's (2015) findings, those that were provided with *how* and *why* information (the transparency manipulation) were less trusting of the autonomous vehicle, perhaps due to an additional need for cognitive resources to comprehend the vehicle's rationale. Further examination of the manipulation check supports this assumption, with 24% of the total fail rate occurring in the transparency condition.

In extension, this study suggests that a failure to describe an autonomous vehicle's rationale for an omission outcome may have contributed to higher trust ratings. When decision rationale is not provided by a vehicle, users may not assign intent or evaluate the vehicle's rationale against their own preferences as would be expected for a transparent autonomous vehicle (see Spranca, Minsk, & Baron, 1991). Automation transparency can be thought of in terms of its influence on perceived vehicle agency. The transparency condition may have led to higher perceptions of vehicle agency due to the vehicle's capability for reason (Waytz, Heafner, & Epley, 2010), and agency implies that a system is an animate object capable of personal

preference and intent (*transparency* manipulation) (see, for example, Bandura, 2006), rather than an impersonal machine operating from a set of opaque, binary programs (*no- transparency* manipulation) (see Pacherie, 2008). That is, when an autonomous vehicle provides information about *how* and *why* it will respond to a moral decision, participants may be less approving of actions due to a higher degree of intentionality inherent in the transparency manipulation.

The positive correlation between trust and blame may indicate a connection between vehicle action and vehicle responsibility. Specifically, because the passenger is not responsible for the vehicle's decision, the passenger is fully capable of trusting the vehicle while also blaming the vehicle for the damage that occurred, a type of moral scapegoating. Rothschild, Landau, Sullivan, and Keefer (2012), for example, found that people tend to preserve their moral identity by relegating responsibility for harmful actions onto a third party. In this thesis the passenger did not have to make a moral decision (i.e., swerve or not swerve), indicating that they trust the vehicle's decision while simultaneously blaming the vehicle for the harm committed.

Lastly, it was expected that when the same moral outcome was performed by a human and an autonomous vehicle, the autonomous vehicle would receive higher acceptance scores. The results suggest that humans may hold machines and humans to different moral standards depending on the moral outcome and the agent responsible for the outcome. Specifically, moral acceptance was higher for autonomous vehicles that made the same utilitarian decision as humans. When, however, autonomous vehicles and humans made non-utilitarian decisions, acceptance scores did not differ. In light of the doctrine of Doing and Allowing, the findings suggest that when active harm needs to be taken (commission bias), it is more morally acceptable for an autonomous vehicle to make the moral decision than a human. In contrast, and unique to this study, when a passive, non-utilitarian decision was made, participants did not hold humans

or autonomous vehicles to different moral standards. This finding may suggest that when passive harm is taking place, it makes no difference who is behind the wheel.

In the case of fully autonomous vehicles, scenarios may arise in which philosophy and human factors collide at the level of system design and personal ethics, in which the Trolley Problem can help explore the effects of automation transparency and ethical outcomes on user trust and blame. Because autonomous vehicles will make decisions on behalf of the driver, it may be important for the driver to be informed of how and why an action will be executed, especially when such actions are morally charged. In fact, Greene and colleagues (2016, p. 4147) stated that: "humans would accept and trust more machines that behave as ethically as other humans in the same environment."

Limitations

The larger body of research suggests that transparency manipulations hinge on the presence (transparent) or absence (non-transparent) of information, and operational definitions of transparency are lacking in several regards. Presence of information about a system's performance-state, for example, does not imply that users will draw correct conclusions about what the system is meaning to convey. When a system provides information to a user, it showcases its potential for performing an action—but determining why an action is being performed may be cognitively demanding and lost in translation on the user depending on the message's content (Koo et al., 2016).

The transparency manipulation employed in this thesis operated on the presence or absence of information paradigms currently used in the literature (Beller, Heesen, &Vollrath, 2013; Kim & Hinds, 2006; Verberne, Ham, & Midden, 2012), when transparency may not be manipulated as casually. More recently, presence of information has been parsed into several

constituent parts, such as *why*, *how,* and a combination of *why* and *how* a system intends to perform an action (Koo et al., 2016), and starker contrasts between each construct may better elicit differences in trust and blame. The manipulation check in this experiment could have captured additional information about the transparency manipulation if participants rated each constituent construct along a continuum.

After reflection, the scenarios in this thesis always resulted in the vehicle swerving into one pedestrian to avoid five pedestrians, and never vice versa. If the vehicle had to swerve to kill five pedestrians, subsequently saving the one pedestrian, the automaticity hypothesis would suggest that the outcome makes no difference because active harm is required. Scenarios that include both swerve combinations will better assess if traditional consequential beliefs bend in the presence of omission and commission biases, while further isolating the effects of transparency and moral outcomes on trust and blame. Measuring participants' moral preferences prior to manipulating the moral outcome would also provide additional experimental utility. It should be noted that it is difficult to draw specific conclusions when moral violations were not specifically measured or balanced in this study. Though a utilitarian outcome seems like an optimal moral decision, the influence of active and passive harm may interfere with accepting such an outcome. Thus, creating instances in which the moral outcome violates or agrees with participants' moral preferences is needed, especially considering how perceived intentional violations can influence blame assignment (Shaver, 1985).

Future Research

This study can be extended by creating moral scenarios for a driving simulator. A driving simulator would enable additional behavioral data to be collected, such as eye movements and brake response times—two variables that could provide further detail about higher order moral

processing. Eye tacking helps identify where attentional resources are distributed and would be a useful method for determining if drivers' eye movements can be mapped to moral dispositions. For example, if a driver adheres to consequentialist beliefs, attentional resources may only be reserved for making utilitarian decisions, as indicated by fixations directed toward one person instead of five leading up to a moral outcome. When paired with a transparent autonomous vehicle, it is plausible that fixations will be directed towards the vehicle's upcoming decision— or redistributed to unrelated driving tasks— so long as the driver perceives the vehicle's algorithm to be in concert with his or her moral preference. From this perspective, incongruence between the driver's moral preference and the vehicle's rationale may be reflected in the driver's fixations, scan patterns (i.e., vertical and horizontal standard deviations), and manual interventions. That is, drivers that disagree with the vehicle's rationale may attempt to intervene by depressing the brake pedal or fixating less frequently on the outcome in protest to the vehicle's decision.

Including different levels of automation, ranging from no automation (manual control), intermediate automation (i.e., ACC), and full automation, would also inform the field's knowledge of moral human-automation interaction. When automation leverages human performance instead of fully replacing human performance, a greater distribution of responsibility across the human-automation partnership is expected to follow. When two parties are equally responsible for moral outcomes, trust and blame may vary as a function of responsibility, adding to the overall conversation of legal liability.

Conclusion

The current study examined the effects of automation transparency and moral outcome on user trust and blame towards fully autonomous vehicles. The results suggest that participants are

more trusting of vehicles that make non-utilitarian decisions without explanations than utilitarian decisions without explanations. The results also found no significant effects of moral outcome or transparency on blame, leading to speculation about the saliency of the transparency manipulation and the importance of measuring moral expectations prior to manipulating moral outcomes. Moral acceptance ratings favored autonomous vehicles that made utilitarian decisions more than humans that made the same decision, but the same effect was not found for non-utilitarian decisions. These findings suggest that moral norms may differ depending on who is performing the action (or inaction). The current research can offer insight regarding the nature of ethical dilemmas on public roadways and carve out a path for future research to explore the ramifications of automated moral decisions.

REFERENCES

Adrian, J., Postal, V., Moessinger, M., Rascle, N., & Charles, A. (2011). Personality traits and executive functions related to on-road driving performance among older drivers. *Accident Analysis & Prevention*, *43*, 1652-1659. doi:10.1016/j.aap.2011.03.023

Ajzen, I., & Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior.* Upper Saddle River, NJ: Prentice Hall.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Aquinas, T. (1988). On law, morality, and politics (W.P. Baumgarth, Ed., & R.J. Regan, Ed. & Trans.). Indianapolis, IN: Hackett.

Bainbridge, L. (1983). Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems 1982* (pp. 129-135).

Bandura, A. (2006). Toward a psychology of human agency. Perspectives on Psychological Science, 1, 164–180. doi: 10.1111/j.1745-6916.2006.00011.x

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74-85.

Barber, B. (1983). The logic and limits of trust.

Bartels, D. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. Cognition, 108, 381–417.

Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the Driver- Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors: The Journal of the Human Factors and Ergonomics So- ciety*, *55*(6), 1130–1141. doi:10.1177/0018720813482327

Bellis, E., & Page, J. (2008). *National motor vehicle crash causation survey (NMVCCS) SAS analytical user's manual* (No. HS-811 053).

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., & Spicer, R. (2002). The economic impact of motor vehicle crashes, 2000. *DOT HS*, *809*, 446.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: are we ready for utilitarian cars?. *arXiv preprint arXiv:1510.03346*.

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573-1576. doi:10.1126/science.aaf2654.

Brown, B. (2018). Tesla says driver ignored warnings from autopilot in fatal California crash. Retrieved from https://www.digitaltrends.com/cars/tesla-autopilot-fatal-crash-warnings-ignored/

Chen, J. Y., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE transactions on systems, man and cybernetics, part C: applications and reviews*, *41*(4), 435-454.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (No. ARL-TR-6905). ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING DIRECTORATE.

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social neuroscience*, *7*(3), 269-279.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*, 10821089. doi: 10.1111/j.1467-9280.2006.01834.x

Coeckelbergh, M. (2016). Responsibility and the Moral Phenomenology of Using Self-Driving Cars. *Applied Artificial Intelligence*, *30*(8), 748-757.

Endsley, M. R., Bolstad, C. A., Jones, D. G., & Riley, J. M. (2003, October). Situation awareness oriented design: from user's cognitive requirements to creating effective supporting technologies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 47, No. 3, pp. 268-272). Sage CA: Los Angeles, CA: SAGE Publications.

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human factors*, *59*(1), 5-27.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77,*167-181. doi.org/10.1016/j.tra.2015.04.003

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Foot, P. (1978). *The problem of abortion and the doctrine of double effect.* In virtues and vices. Oxford: Blackwell.

Forster, Y., Naujoks, F., & Neukum, A. (2017). Increasing anthropomorphism and trust in automated driving functions by adding speech output. In *2017 IEEE Intelligent Vehicles Symposium (IV)* (pp. 365–372). IEEE. https://doi.org/10.1109/IVS.2017.7995746

Friedman, B., & Millett, L. I. (1997). *Reasoning about computers as moral agents: A research note* (p. 205). Stanford: CSLI Publications.

Glass, A., McGuinness, D. L., & Wolverton, M. (2008, January). Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces* (pp. 227-236). ACM.

Gleicher, F., Kost, K. A., Baker, S. M., Strathman, A. J., Richman, S. A., & Sherman, S. J. (1990). The role of counterfactual thinking in judgments of affect. *Personality and Social Psychology Bulletin*, *16*(2), 284-295.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In *Road vehicle automation* (pp. 93-102). Springer International Publishing.

Gravetter, F., & Wallnau, L. (2014). Essentials of statistics for the behavioral sciences (8th ed.). Belmont, CA: Wadsworth.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *science*, *315*(5812), 619-619.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364-371. doi: 10.1016/j.cognition.2009.02.001

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144-1154. doi.org/10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389-400. doi.org/10.1016/j.neuron.2004.09.027

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. C. (2016, February). Embedding Ethical Principles in Collective Decision Support Systems. In *AAAI* (pp. 4147-4151).

Hauser, M., Cushman, F., Young, L., Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. Mind & Language, 22, 1–21.

Hennessy, D. A., & Wiesenthal, D. L. (1999). Traffic congestion, driver stress, and driver aggression. *Aggressive behavior*, *25*, 409-423. doi:10.1002/(SICI)1098-2337(1999)25:6<409::AID-AB2>3.0.CO;2-0

Hoff, K. A., Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust . Human Factors, 57, 407–434.

Jenness, J. W., Lerner N. D., Mazor, S., Osberg, J. S., & Tefft, B. C. (2008). Use of advanced in-vehicle technology by young and older early adopters. Survey results on adaptive cruise control systems. Report no. DOT HS 810 917. Washington, DC: US Department of Transportation, National Highway Traffic Safety Administration.

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*, 53-71. doi.org/10.1207/S15327566IJCE0401_04

Kiernan, K. (2015). Human Factors Considerations in Autonomous Lethal Unmanned Aerial Systems. *Aviation / Aeronautics / Aerospace International Research Conference*.

Kim, T., & Hinds, P. (2006, September). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on* (pp. 80-85). IEEE.

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

Körber, M., Prasch, L., & Bengler, K. (2018). Why do I have to drive now? Post hoc explanations of takeover requests. *Human factors*, *60*(3), 305-323.

Kumfer, W. J., Levulis, S. J., Olson, M. D., & Burgess, R. A. (2016, September). A Human Factors Perspective on Ethical Concerns of Vehicle Automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1844-1848). Sage CA: Los Angeles, CA: SAGE Publications.

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

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, *40*, 153-184. doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*, 50-80. doi: 10.1518/hfes.46.1.50_3039

Lee, J. D. (2012). Trust, trustworthiness, and trustability. *Presentation at the Workshop on Human Machine Trust for Robust Autonomous Systems*. Ocala, FL.

Lyons, J. B. (2013, March). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., ... & Shively, R. (2017). Shaping trust through transparent design: theoretical and experimental guidelines. In *Advances in Human Factors in Robots and Unmanned Systems* (pp. 127-136). Springer, Cham.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In J. Forgas, K. Fiedler, & C. Sedikides (Eds.), Social thinking and interpersonal behaviour (pp. 311–329). Philadelphia, PA: Psychology Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Malle, B. F., Monroe, A. E., & Guglielmo, S. (2014). Paths to blame and paths to convergence. *Psychological Inquiry*, *25*(2), 251-260.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124). ACM.

Maxwell, S. E., & Delaney, H. D. (2003). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors*, *58*(3), 401-415.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194-210.

Milgram, S. 1974. Obedience to authority. New York: Harper & Row.

Mill, J. S. (1901). *Utilitarianism*. Longmans, Green and Company.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology*, *8*, 47-63. doi.org/10.1207/s15327108ijap0801_3

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological science*, *19*(6), 549-557.

Muir, B. M., & Moray, N. (1996). Trust in automation: 2. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*, 429–460. doi.org/10.1080/00140139608964474

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates?. *International Journal of Human-Computer Studies*, *45*(6), 669-678.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied*, *7*(3), 171.

National Highway Traffic Safety Administration. (2015). The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised). *Annals of Emergency Medicine*, *66*, 194-196.

National Highway Traffic Safety Administration, Department of Transportation (2017). *Automated driving systems 2.0: A vision for safety.* Retrieved from https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf

National Safety Council. (2017). Motor Vehicle Deaths in 2016 Estimated to be Highest in Nine Years. Retrieved from http://www.nsc.org/Connect/NSCNewsReleases/Lists/Posts/Post.aspx?ID=180

Nothdurft, F., & Minker, W. (2016). Justification and transparency explanations in dialogue systems to maintain human-computer trust. In *Situated Dialog in Speech-Based Human-Computer Interaction* (pp. 41-50). Springer International Publishing.

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem. *Ethical Theory Moral Pract*, *19*(5), 1275-1289.

Oduor, K. F., & Wiebe, E. N. (2008, September). The effects of automated decision algorithm modality and transparency on reported trust and task performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, No. 4, pp. 302-306). Sage CA: Los Angeles, CA: SAGE Publications.

Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. (2014, June). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned Systems Technology XVI* (Vol. 9084, p. 90840E). International Society for Optics and Photonics.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179-217.

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

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

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of doing and allowing. *The Philosophical Review*, *98*(3), 287-312.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

Rickless, S.C. (1997). The Doctrine of doing and allowing. The Philosophical Review, 106(4), 555-575.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making*, *3*(4), 263-277.

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

Ross, V., Jongen, E., Brijs, T., Ruiter, R., Brijs, K., & Wets, G. (2015). The relation between cognitive control and risky driving in young novice drivers. *Applied Neuropsychology: Adult*, *22*, 61-72. doi.org/10.1080/23279095.2013.838958

Rothschild, Z. K., Landau, M. J., Sullivan, D., & Keefer, L. A. (2012). A dual-motive model of scapegoating: Displacing blame to reduce guilt or increase control. *Journal of Personality and Social Psychology*, *102*(6), 1148.

SAE International. (2016). *SAE surface vehicle recommended practice report: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles* (No. J3016). Warrendale, PA: Author.

Sanders, T. L., Wixon, T., Schafer, K. E., Chen, J. Y., & Hancock, P. A. (2014, March). The influence of modality and transparency on trust in human-robot interaction. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on* (pp. 156-159). IEEE.

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.

Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *International journal of human-computer studies*, *65*(3), 192-205.

Shaver, K. G. (2012). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Science & Business Media.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*, (Tech. Rep.). Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA.

Sheridan, T. B., & Hennessy, R. T. (1984). *Research and modeling of supervisory control behavior.* Washington, D.C.: National Academy Press.

Singh, S. (2015). *Critical reasons for crashes investigated in the national motor vehicle crash causation survey* (No. DOT HS 812 115).

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making?. *International Journal of Human-Computer Studies*, *51*, 991-1006. doi.org/10.1006/ijhc.1999.0252

Spranca, M., Minsk, E., Baron, J. (1991). Omission and commission in judgment and choice. Journal of Experimental Social Psychology, 27(1), 76–105.

Stanton, N.A., Marsden, P., 1996. From fly-by-wire to drive-by-wire: safety implications of automation in vehicles. Safety Science 24 (1), 35–49.

Stokols, D., Novaco, R. W., Stokols, J., & Campbell, J. (1978). Traffic congestion, Type A behavior, and stress. *Journal of Applied Psychology*, *63*, 467. doi.org/10.1037/0021-9010.63.4.467

Triplett, T., Santos, R., & Rosenbloom, S. (2015). *American Driving Survey: Methodology and Year One Results,* (Research Report No. 01561011). Washington, DC: AAA Foundation Traffic Safety.

Verberne, F. M., Ham, J., & Midden, C. J. (2012). Trust in smart systems sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 799-810. doi: 10.1177/0018720812443825

Verberne, F., Ham, J., & Midden, C. (2012, June). Trusting Automation Technology for Safer Roads: The Effect of Shared Driving Goals. In *Persuasive Technology: Design for Health and Safety;*

The 7th International Conference on Persuasive Technology; Persuasive 2012; Sweden; June 6-8; Adjunct Proceedings (No. 068, pp. 57-60). Linköping University Electronic Press.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, *22*(4), 565-582.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human factors*, *51*(3), 281-291.

Wang, L., Jamieson, G. A., & Hollands, J. G. (2011, September). The effects of design features on users' trust in and reliance on a combat identification system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, No. 1, pp. 375-379). SAGE Publications.

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219-232.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

Wintersberger, P., Frison, A. K., Riener, A., & Boyle, L. N. (2016). Towards a Personalized Trust Model for Highly Automated Driving. *Mensch und Computer 2016–Workshopband*.

World Health Organization. Violence, Injury Prevention, & World Health Organization. (2013). *Global status report on road safety 2013: supporting a decade of action*. World Health Organization.

World Health Organization. (2016). *World Health Statistics 2016: Monitoring Health for the Sustainable Development Goals (SDGs)*. World Health Organization.

Wright, J. L. (2015). Transparency in Human-agent Teaming and its Effect on Automation-induced Complacency. *Procedia Manufacturing*, *3*, 968-973

APPENDIX A

AUTONOMOUS VEHICLE ATTITUDE'S SURVEY

There has been a need to understand how peoples' attitudes towards autonomous vehicles can differ based on the vehicle's automation setting. I want to explore how these attitudes can vary amongst a sample of students.

I am writing to ask for your help with my research effort. As a master's student in Experimental Psychology, I am pursuing this topic as part of my research fundamentals requirement. I am completing my thesis under the supervision of Dr. Bryan Porter, Associate Dean of the Graduate School.

Attached is a brief survey to collect information about user attitudes towards autonomous vehicles and how they may or may not be affected by certain driving scenarios. Survey responses are anonymous. Personally-identifiable data will not be collected. In addition, survey questions will be randomized into block groupings, with each participant receiving one randomly assigned block of questions. Therefore, I will be unable to share any individual's data with his/her program. You will earn 0.5 credits to be used toward your research participation requirement in your psychology course.

If you have questions, do not hesitate to contact me at 757-812-9813 or at my email: nhatf001@odu.edu.

Thank you very much for considering my request. I look forward to hearing from you should the need arise.

Sincerely,

Nathan Hatfield| nhatf001@odu.edu
Researcher | Masters Candidate, Experimental Psychology
Old Dominion University
Norfolk, Virginia 23259
757-812-9813

Supervisor:
Bryan E. Porter, Ph.D. | bporter@odu.edu
Associate Dean, the Graduate School
Old Dominion University
Norfolk, Virginia 23259
757-683-3259

AUTONOMOUS VEHICLE ATTITUDE SURVEY

Directions (Please read carefully)

Thank you for participating in our survey. The data we collect from you will be anonymous and completely confidential, and will be used to support data collection efforts to satisfy thesis requirements.

On the following page you will be presented with a description of a situation and an action that an autonomous vehicle in that situation might perform in response to that situation. Your job is to tell us (1) whether you think it would be morally acceptable for the autonomous vehicle to perform this action, (2) how morally acceptable/ unacceptable this action would be, (3) followed by questions to gauge your attitude towards the vehicle based on its decision.

The questions concern the action's moral acceptability, and not what yourself or anyone else would actually do in the situation described.

You might feel that the situation as we describe it is not realistic. For example, it might say that if the autonomous vehicle does X, then Y will happen, and you might think that this is not realistic, that Y might not necessarily happen if the autonomous vehicle does X. If you find yourself having these sorts of doubts, "suspend disbelief" just as you would at an unrealistic movie and assume that this situation really is the way it's described.

Likewise, you may feel that you need more information than is provided about the situation before you can give your answer. If this happens, you should make your best guess about what you think the situation is like without making any unnecessary assumptions.

Do you have any questions? If so, please ask the experimenter by emailing nhatf001@odu.edu. Otherwise, please proceed to the next page.

Thank you in advance for your time, interest, and support.

Nathan Hatfield
Nhatf001@odu.edu
757-812-9813
Old Dominion University
Norfolk, Virginia

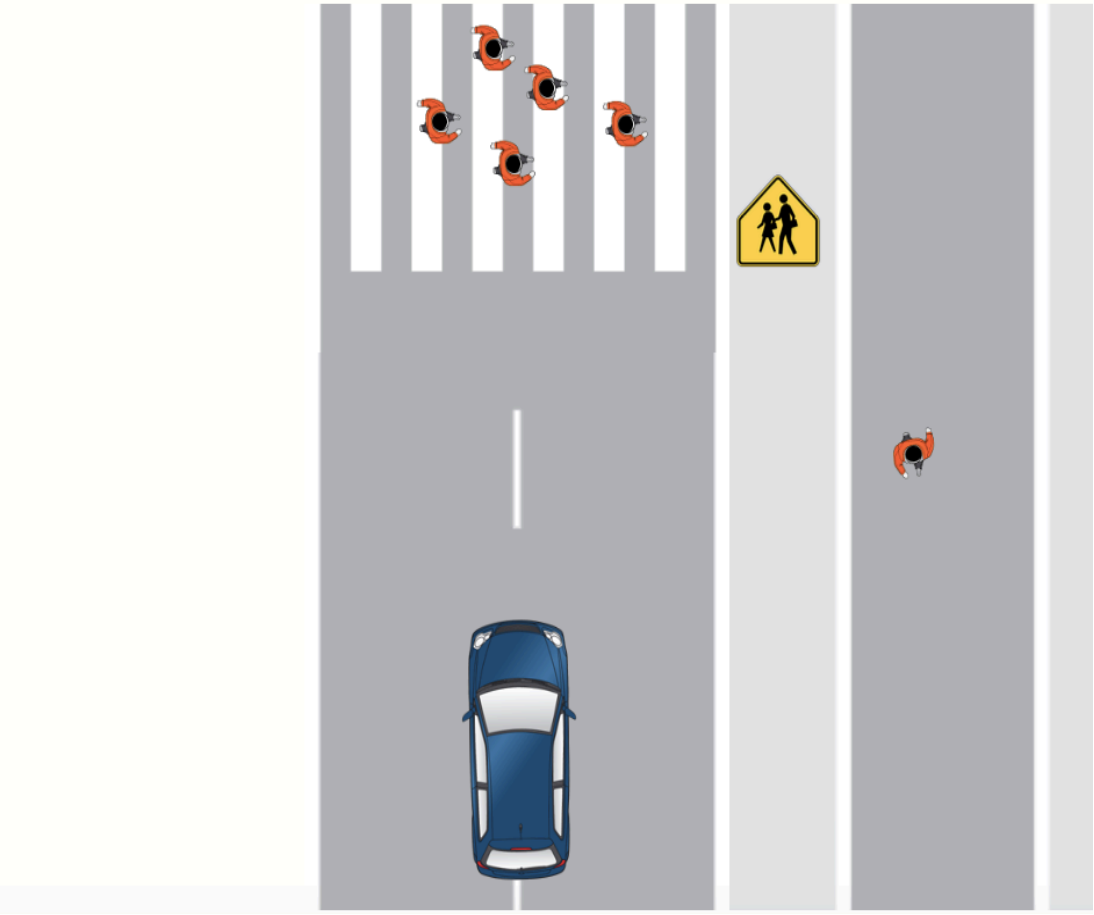
No Transparency Scenario w/ Utilitarian Outcome

You have been assigned to a random block of questions as related to autonomous vehicles. Please read the following scenario and provide a response to the questions or items below:

On the roadway is a self-driving car. The self-driving car is capable of performing all driving functions under all conditions. The self-driving car can change lanes, turn, use signals, break, accelerate, and monitor the roadway without the need for human intervention, which means that the occupant does not have any control of the vehicle.

A passenger is riding in a self-driving car that is quickly approaching an intersection. In the middle of the intersection is a group of five pedestrians legally crossing the road. On the sidewalk, however, is a single pedestrian.

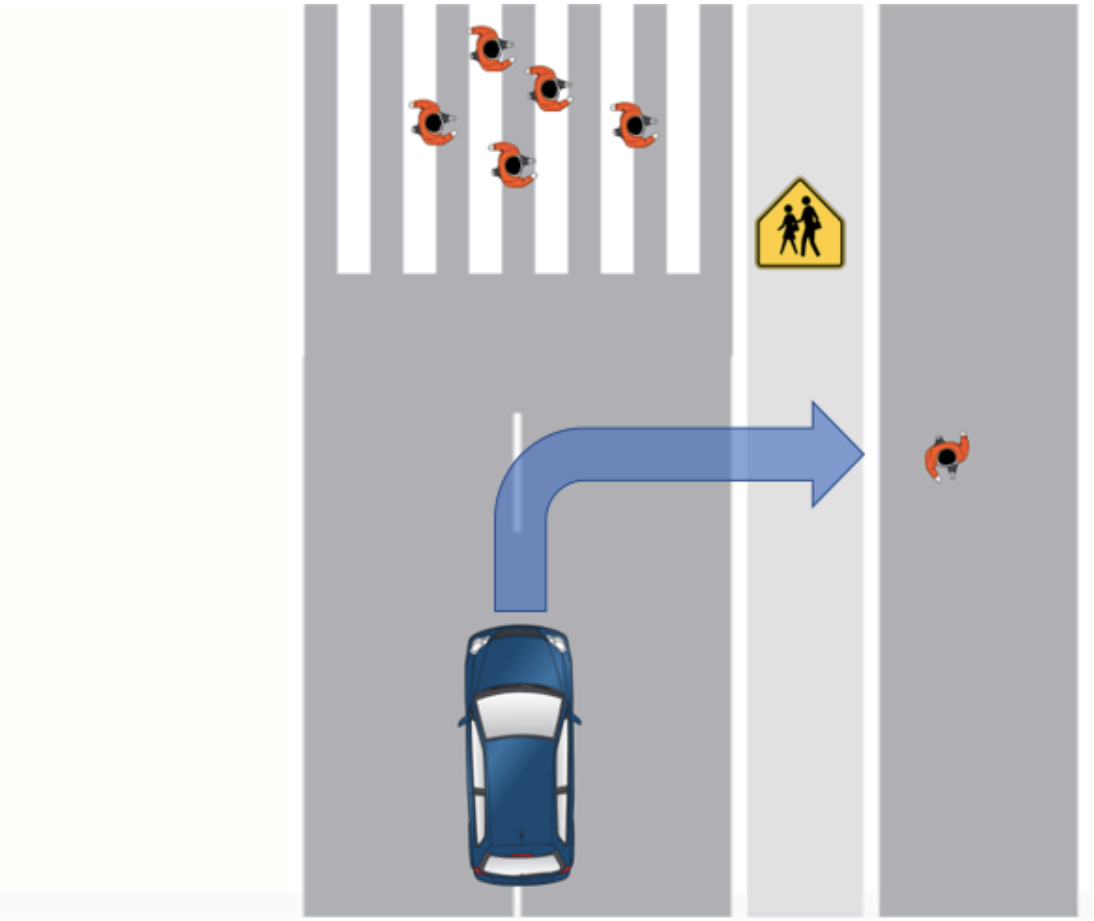
The self-driving car is faced with two options: continue straight through the intersection where it will kill 5 pedestrians, or swerve to the right where it will kill one pedestrian.



Is it morally acceptable for the self-driving car to direct itself toward the single pedestrian?

- Please circle one answer: YES / NO
- To what extent is this action morally acceptable?
    - (Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

The self-driving car directed itself toward the single pedestrian.



- **Blame scale (Scale titles will not be placed in document that participants see)**
    - Example question: “How much is the self-driving car itself at fault?”

Now imagine that this accident had occurred in the real world:

- **Second part of blame scale**

Based on what you know about the self-driving car’s capabilities and performance, please answer the following questions:

- **Trust scale (Scale titles will not be placed in document that participants see)**
    - Example question**: “**The autonomous vehicle is deceptive.”
    - **Manipulation check**: did the self-driving car provide you with an explanation for why it made the decision it did? Y/N

Now imagine that a human driver is in control of the vehicle, recognizes the same facts, and faces the same decision.

"Is it morally acceptable for the driver to direct the car toward the single pedestrian?"

- Please circle one answer: YES / NO

(Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

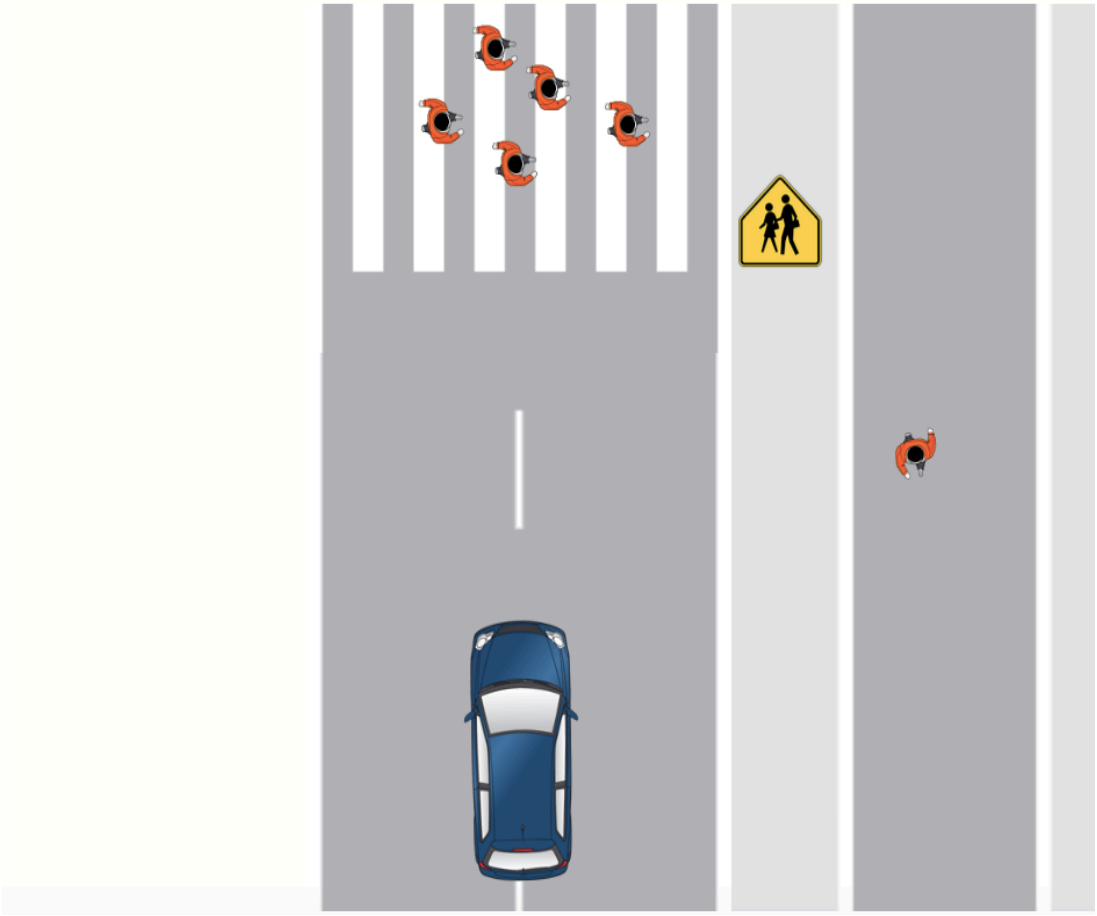
No Transparency Scenario w/ Non Utilitarian Outcome

You have been assigned to a random block of questions as related to autonomous vehicles. Please read the following scenario and provide a response to the questions or items below:

On the roadway is a self-driving car. The self-driving car is capable of performing all driving functions under all conditions. The self-driving car can change lanes, turn, use signals, break, accelerate, and monitor the roadway without the need for human intervention, which means that the occupant does not have any control of the vehicle.

A passenger is riding in self-driving car that is quickly approaching an intersection. In the middle of the intersection is a group of five pedestrians legally crossing the road. On the sidewalk, however, is a single pedestrian.

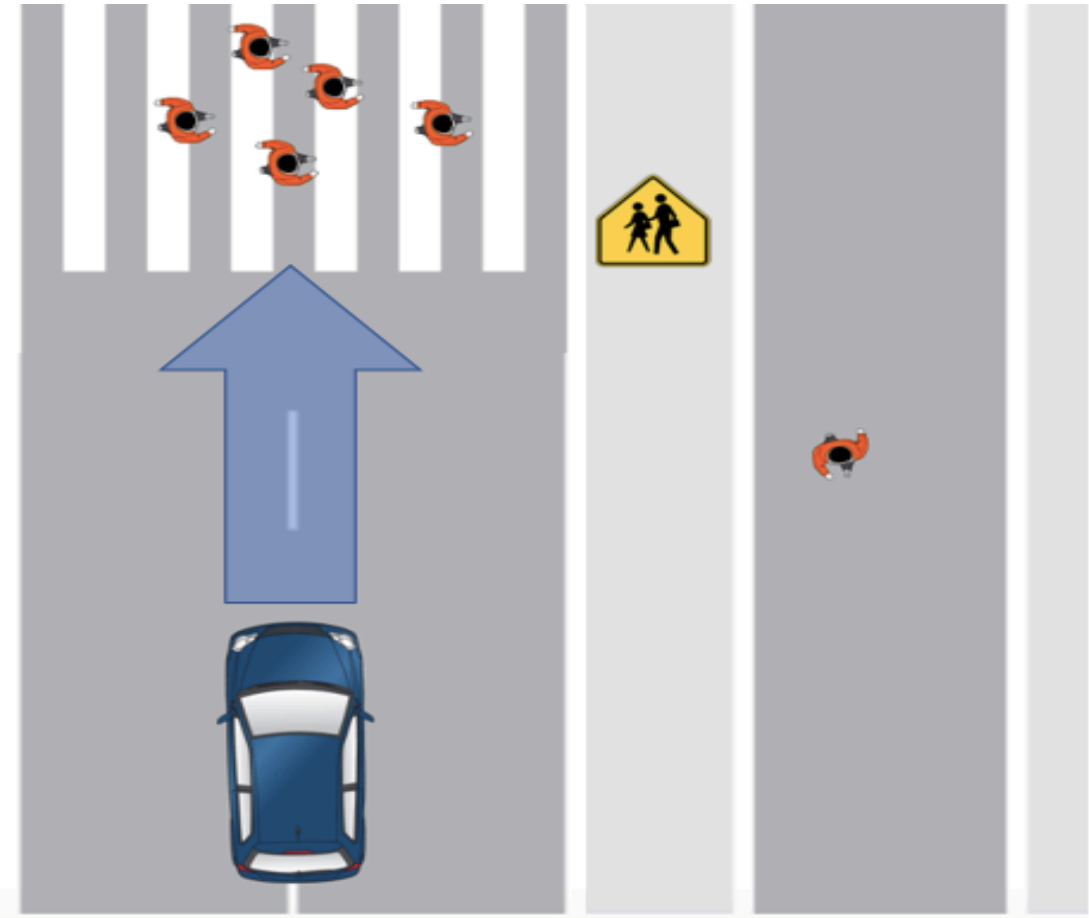
The self-driving car is faced with two options: continue straight through the intersection where it will kill 5 pedestrians, or swerve to the right where it will kill one pedestrian.



Is it morally acceptable for the self-driving car to direct itself toward the single pedestrian?

- Please circle one answer: YES / NO
- To what extent is this action morally acceptable?
    - (Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

The self-driving car directed itself towards the five pedestrians.



- **Blame scale (Scale titles will not be placed in document that participants see)**

Now imagine that this accident had occurred in the real world:

- **Second part of blame scale**

Based on what you know about the self-driving car's capabilities and performance, please answer the following questions:

- **Trust scale (Scale titles will not be placed in document that participants see)**

- **Manipulation check**: did the self-driving car provide you with an explanation for why it made the decision it did? Y/N

Now imagine that a human driver is in control of the vehicle, recognizes the same facts, and faces the same decision.

"Is it morally acceptable for the driver to not direct the car toward the single pedestrian?"

- Please circle one answer: YES / NO

(Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

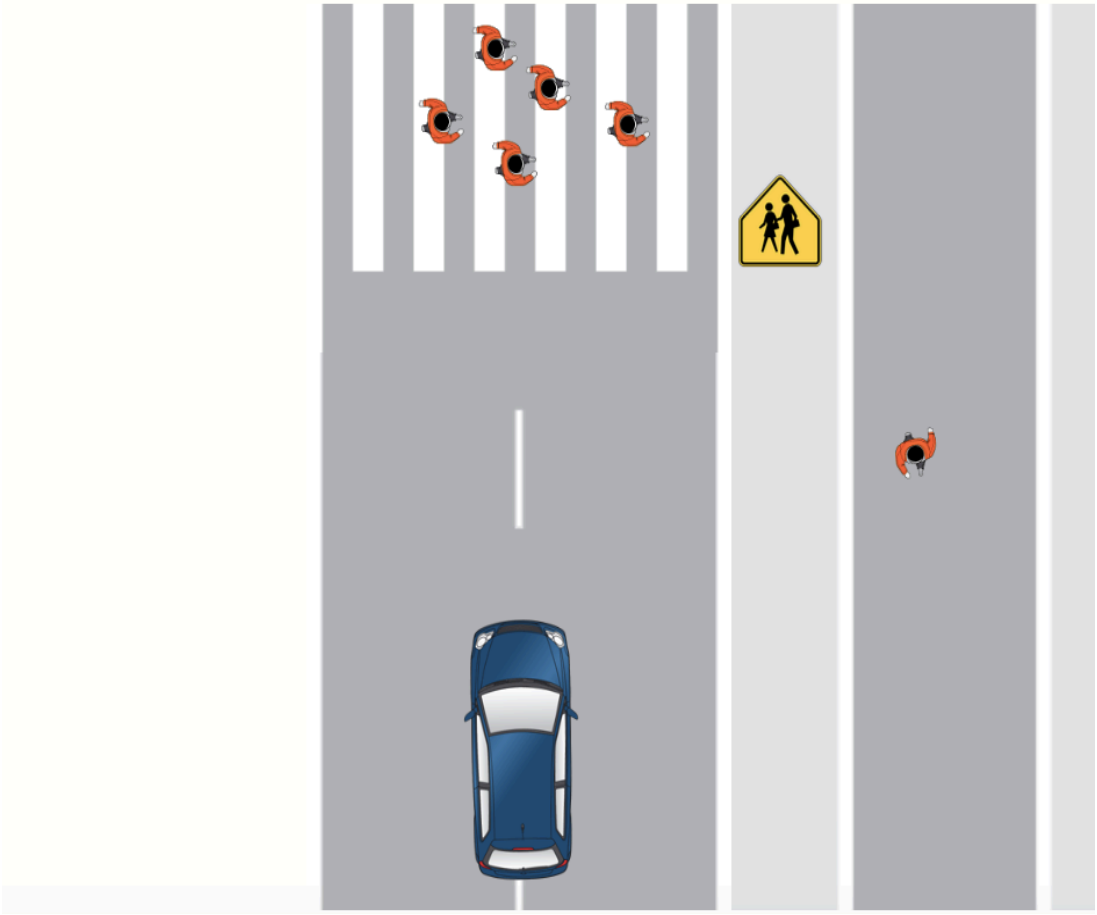
High Transparency Scenario w/ Utilitarian Outcome

You have been assigned to a random block of questions as related to autonomous vehicles. Please read the following scenario and provide a response to the questions or items below:

On the roadway is a self-driving car. The self-driving car is capable of performing all driving functions under all conditions. The self-driving car can change lanes, turn, use signals, break, accelerate, and monitor the roadway without the need for human intervention, which means that the occupant does not have any control of the vehicle.

A passenger is riding in self-driving car that is quickly approaching an intersection. In the middle of the intersection is a group of five pedestrians legally crossing the road. On the sidewalk, however, is a single pedestrian.

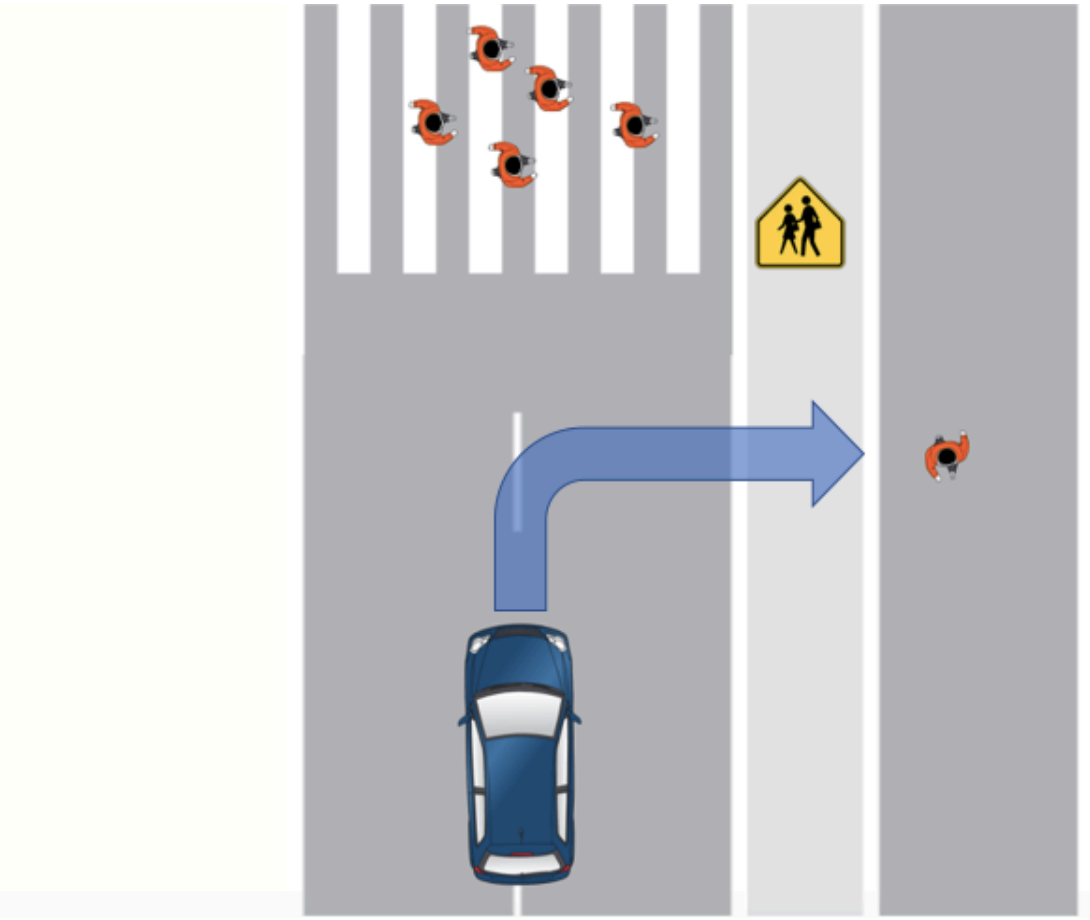
The self-driving car is faced with two options: continue straight through the intersection where it will kill 5 pedestrians, or swerve to the right where it will kill one pedestrian.



Is it morally acceptable for the self-driving car toward the single pedestrian?

- Please circle one answer: YES / NO
- To what extent is this action morally acceptable?
    - (Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

Before the self-driving car acted, it informed the occupant that an ensuing crash was getting ready to occur. This information was provided to the occupant via an icon warning on the self-driving car's dashboard. In addition to the icon warning, the self-driving car further notified the passenger that it would direct itself toward the single pedestrian to ensure the least amount of damage occurred.



- **Blame scale (Scale titles will not be placed in document that participants see)**

Now imagine that this accident had occurred in the real world:

- **Second part of blame scale**

Based on what you know about the self-driving car's capabilities and performance, please answer the following questions:

- **Trust scale (Scale titles will not be placed in document that participants see)**

- **Manipulation check**: did the self-driving car provide you with an explanation for why it made the decision it did? Y/N

Now imagine that a human driver is in control of the vehicle, recognizes the same facts, and faces the same decision.

"Is it morally acceptable for the driver to direct the car toward the single pedestrian?"

- Please circle one answer: YES / NO

(Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

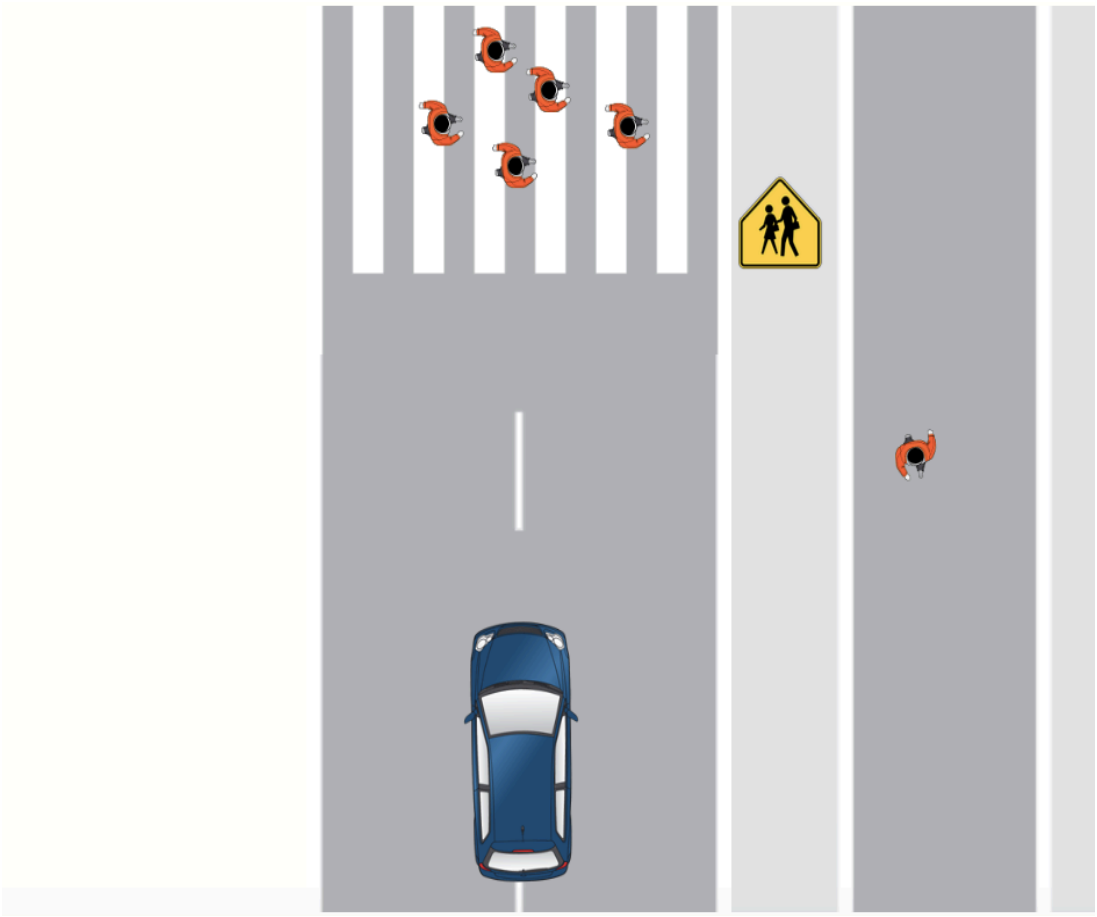
High Transparency Scenario w/ Non-Utilitarian Outcome

You have been assigned to a random block of questions as related to autonomous vehicles. Please read the following scenario and provide a response to the questions or items below:

On the roadway is a self-driving car. The self-driving car is capable of performing all driving functions under all conditions. The self-driving car can change lanes, turn, use signals, break, accelerate, and monitor the roadway without the need for human intervention, which means that the occupant does not have any control of the vehicle.

The self-driving car is quickly approaching an intersection. In the middle of the intersection is a group of five pedestrians legally crossing the road. On the sidewalk, however, is a single pedestrian.

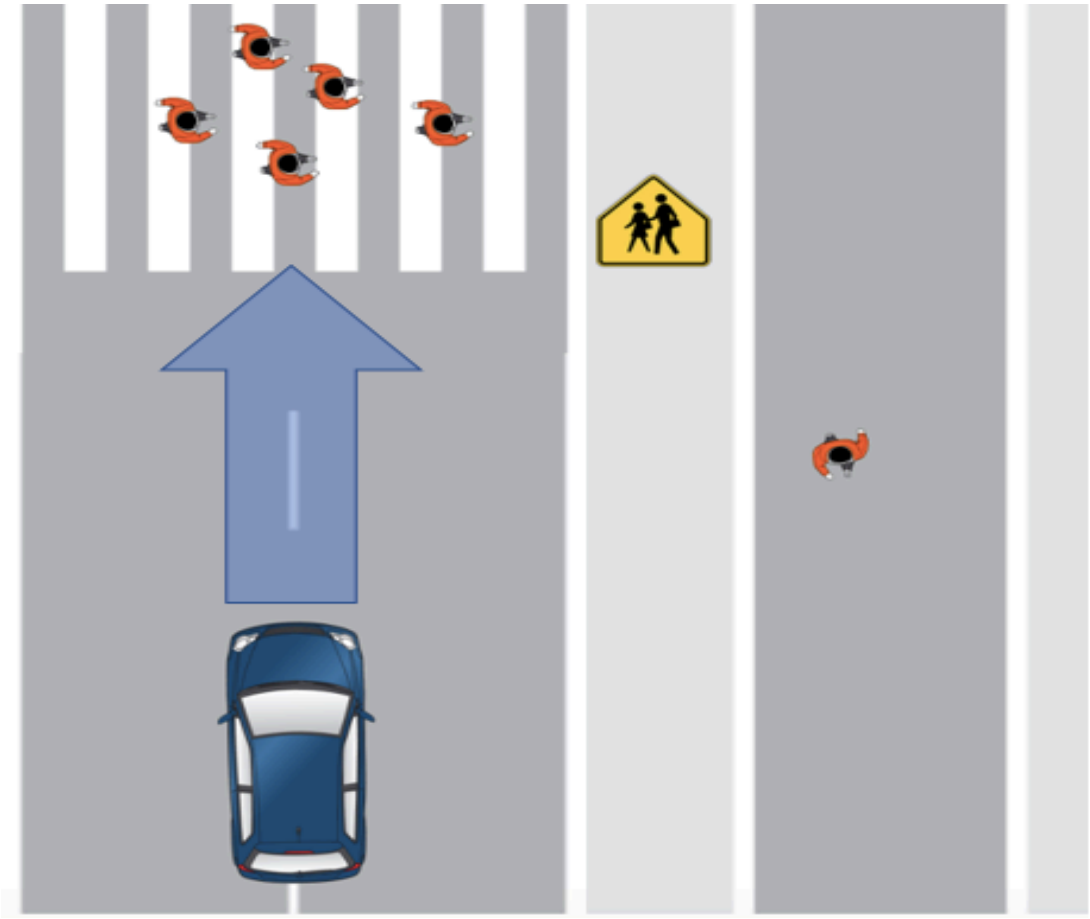
The self-driving car is faced with two options: continue straight through the intersection where it will kill 5 pedestrians, or swerve to the right where it will kill one pedestrian.



Is it morally acceptable for the self-driving car to direct itself toward the single pedestrian?

- Please circle one answer: YES / NO
- To what extent is this action morally acceptable?
    - (Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

Before the self-driving car acted, it informed the occupant that an ensuing crash was getting ready to occur. This information was provided to the occupant via an icon warning on the self-driving car's dashboard. In addition to the icon warning, the self-driving car notified the passenger that it would direct itself toward the five pedestrians to ensure the least amount of damage occurred.



- **Blame scale (Scale titles will not be placed in document that participants see)**

Now imagine that this accident had occurred in the real world:

- **Second part of blame scale**

Based on what you know about the self-driving car's capabilities and performance, please answer the following questions:

- **Trust scale (Scale titles will not be placed in document that participants see)**

- **Manipulation check**: did the self-driving car provide you with an explanation for why it made the decision it did? Y/N

Now imagine that a human driver is in control of the vehicle, recognizes the same facts, and faces the same decision.

"Is it morally acceptable for the driver to not direct the car toward the single pedestrian?"

- Please circle one answer: YES / NO
    - (Completely unacceptable) 1 2 3 4 5 6 7 8 9 (Completely acceptable)

APPENDIX B

TRUST IN AUTOMATION SCALE

1. The autonomous vehicle is deceptive
2. The autonomous vehicle behaves in an underhanded manner
3. I am suspicious of the autonomous vehicle's intent, action or outputs.
4. I am wary of the autonomous vehicle.
5. The autonomous vehicle will have a harmful or injurious outcome.
6. I would be confident in the autonomous vehicle if it were my own.
7. The autonomous vehicle provides security.
8. The autonomous vehicle has integrity
9. The autonomous vehicle is dependable.
10. The autonomous vehicle is reliable.
11. I would trust the autonomous vehicle if it were my own.
12. I am familiar with the autonomous vehicle's system

APPENDIX C

VEHICLE BLAME SCALE

Please consider the accident that just occurred and answer these questions about responsibility.

Not at All----------Very Much

1	How much is the passenger at fault?	0----1----2----3----4----5----6----7----8----9----10
2	How much is the self-driving car itself at fault?	0----1----2----3----4----5----6----7----8----9----10
3	How much are the people who designed the self-driving car at fault?	0----1----2----3----4----5----6----7----8----9----10
4	How much is the company that developed the self-driving car at fault?	0----1----2----3----4----5----6----7----8----9----10

Now imagine that this accident had occurred in the real world, with another driver driving the car.

Not at All----------Very Much

1	How strongly do you feel that the passenger of the car should be sent to jail?	0----1----2----3----4----5----6----7----8----9----10
2	How strongly do you feel that the self-driving car should be destroyed?	0----1----2----3----4----5----6----7----8----9----10
3	How strongly do you feel that the engineer who designed the self-driving car should be punished?	0----1----2----3----4----5----6----7----8----9----10
4	How strongly do you feel that the company who designed the self-driving car should be punished financially?	0----1----2----3----4----5----6----7----8----9----10

Note: Items 2-4 in first and second section constitute the composite for blame toward the car for the accident.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113-117.

APPENDIX D

DEMOGRAPHIC QUESTIONS

Please indicate you gender:

- Female
- Genderqueer/Gender Non-conforming
- Male
- Transgender Male
- Transgender Female
- Preferred identity (in addition to or not listed above):_____
- Prefer not to state

Are you of Hispanic, Latino, or Spanish origin?

- Yes
- No
- Prefer not to state

Please indicate your race/ ethnicity:

- White
- Black or African American
- American Indian or Alaska Native
- Asian Indian
- Chinese
- Filipino
- Japanese
- Korean
- Vietnamese
- Native Hawaiian
- Guamanian or Chamorro
- Samoan
- Other Pacific Islander
- Other Asian
- Other:_____
- Prefer not to state

Citizenship:

- US
- Non-US(please specify home country):_____

Do you commute to ODU? (Choices) Yes, No

(Follow up) If so, how long is your average commute in minutes? (Fill in) ___

How many years have you had your driver's license? Please put 0 for less than a year, and N/A if you do not have a driver's license. (Fill in)____

How often do you drive your motor vehicle in a week? (Choices) Everyday, 3-5 times a week, Once or twice a week, I rarely drive, I do not drive/I do not have a car

(Follow up) Estimate miles driven per week. (Fill in) ___

Have you ever received a ticket for a driving violation? (Choices) Yes, No

Have you ever been involved in a traffic accident? (Choices) Yes, No

Have you ever had an accident or near-accident due to sleepiness? (Choices) Never, within the last 6 months, within the last year, within the last 5 years

Thank you for completing this survey!

VITA

Nathan Andrew Hatfield
Department of Psychology
Old Dominion University
Norfolk, VA 23529

Education

Bachelor of Science, Psychology, Christopher Newport University,
May 2014

Publications

Hatfield, N., Yamani, Y., Palmer, D. B., Karpinsky-Mosley, N. D., Horrey, W. J., & Samuel, S. Comparison of visual sampling patterns under simulated L2 and L0 systems. Proceedings of the Human Factors and Ergonomics Society 2018 Annual Meeting. Philadelphia, PA. (Extended Abstract)

Yamani, Y., Bicaksiz, P., Palmer, D. B., **Hatfield, N.** & Samuel, S. (2018). Evaluation of the effectiveness of a gaze-based training intervention on latent hazard anticipation skills for young drivers: A driving simulator study. Safety, 4, 18. DOI: 10.3390/safety4020018.

Presentations

Cartwright, K. B., **Hatfield, N. A.,** Marshall, T (2015). Utility of a New Measure of Reading-Specific Executive Skill for Predicting Reading Comprehension: Implications for Practice. Poster presented at the 2015 biennial meeting of the Society for Research in Child Development.

Hatfield, N., Yamani, Y., Palmer, D. B., Karpinsky-Mosley, N. D., Horrey, W. J., & Samuel, S. Comparison of visual sampling patterns under simulated L2 and L0 systems. Presentation given at the Human Factors and Ergonomics Society 2018 Annual Meeting. Philadelphia, PA.