

STARS


University of Central Florida
STARS

Electronic Theses and Dissertations, 2004-2019

2017

The Role of Accounts and Apologies in Mitigating Blame toward Human and Machine Agents

Kimberly Stowers
University of Central Florida

 Part of the [Cognitive Psychology Commons](#)
Find similar works at: <https://stars.library.ucf.edu/etd>
University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Stowers, Kimberly, "The Role of Accounts and Apologies in Mitigating Blame toward Human and Machine Agents" (2017). *Electronic Theses and Dissertations, 2004-2019*. 5932.
<https://stars.library.ucf.edu/etd/5932>



MORAL BLAMEWORTHINESS AND TRUSTWORTHINESS:
THE ROLE OF ACCOUNTS AND APOLOGIES IN PERCEPTIONS OF
HUMAN AND MACHINE AGENTS

by

KIMBERLY STOWERS
B.S. University of Central Florida, 2013
M.S. University of Central Florida, 2015

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Summer Term
2017

Major Professor: Peter A. Hancock

© 2017 Kimberly Stowers

ABSTRACT

Would you trust a machine to make life-or-death decisions about your health and safety? Machines today are capable of achieving much more than they could 30 years ago—and the same will be said for machines that exist 30 years from now. The rise of intelligence in machines has resulted in humans entrusting them with ever-increasing responsibility. With this has arisen the question of whether machines should be given equal responsibility to humans—or if humans will ever perceive machines as being accountable for such responsibility. For example, if an intelligent machine accidentally harms a person, should it be blamed for its mistake? Should it be trusted to continue interacting with humans? Furthermore, how does the assignment of moral blame and trustworthiness toward machines compare to such assignment to humans who harm others? I answer these questions by exploring differences in moral blame and trustworthiness attributed to human and machine agents who make harmful moral mistakes. Additionally, I examine whether the knowledge and type of *reason*, as well as *apology*, for the harmful incident affects perceptions of the parties involved. In order to fill the gaps in understanding between topics in moral psychology, cognitive psychology, and artificial intelligence, valuable information from each of these fields have been combined to guide the research study being presented herein.

For every child who dares to dream.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser, Dr. Peter Hancock, for supporting me in pursuing every endeavor I've dared under his guidance. I would also like to thank my committee members, Dr. Florian Jentsch, Dr. Mustapha Mouloua, Dr. Jessie Chen, and Dr. Daniel Barber, for their guidance and support throughout my doctoral studies.

Thank you to Dr. Valerie Sims for providing expertise on the development of my dissertation topic. Great thanks are due to Olivia Newton for always being available to brainstorm with me, as well as Keith MacArthur for helping me navigate the evolving intricacies of the IRB. Additionally, I would like to thank Dr. Jessica Cruit for her support, and Dolores Rodriguez-Romero for guiding me through the many logistics of completing this program.

I would like to thank Dr. Eduardo Salas, the man who taught me nearly everything I know about how to grow and maintain a robust research program, as well as Dr. Shawn Burke for showing me the true meaning of perseverance and hard work. Additionally, thank you to the colleagues and friends who have supported me these last 4 years—you know who you are!

Thank you to Chaundra and Judy for always loving me without judgment. Thank you to my grandparents and extended family for their love and support throughout my life. Thank you to my amazing husband, Chas, for embarking on this incredible journey with me. You are everything and more. No words exist to describe the gratitude I have for my sister, Emily, who is so deeply a part of me—we are truly one. Last and certainly not least, thank you to my parents, Stephanie and Bryan. As you have been here for me, I will be there for you.

TABLE OF CONTENTS

LIST OF FIGURES	ix
LIST OF TABLES	xi
CHAPTER ONE: INTRODUCTION.....	1
Research Purpose	5
CHAPTER TWO: REVIEW OF LITERATURE.....	6
Agent to Agent	6
Agent Autonomy.....	8
Whose Morals?.....	10
Morality in Humans	11
Morality in Machines.....	14
Playing God.....	18
Equal Blame.....	20
Should I Trust You?	27
“I’m sorry!”—Accounts and Apologies.....	30
Accounts	31
Apologies	33
Either, Or, or Both?.....	35
CHAPTER THREE: PROPOSED EXPERIMENTAL PROCEDURE	38

Design.....	38
Task Overview	38
Materials.....	41
Moral Blame	41
Perceived Trustworthiness	42
Displaced Blame	43
Perceived Agency	44
Participants	44
Recruitment.....	44
Sample.....	45
Procedure.....	46
CHAPTER FOUR: EXPERIMENTAL RESULTS	47
Moral Blame.....	47
Sex Differences	48
Examination of Hypotheses	49
Perceived Trustworthiness	52
Sex Differences	53
Examination of Hypotheses	53
Trustworthiness Dimensions.....	56

Displaced Blame and Agency	59
Displaced Blame	59
Perceived Agency	60
CHAPTER FIVE: DISCUSSION.....	62
How Do I Blame Thee?.....	63
Are You Worthy of My Trust?.....	66
Yes, Sex Matters	69
Theoretical and Practical Implications	70
Theoretical Implications	70
Practical Implications.....	71
Future Research.....	72
APPENDIX A: NARRATIVES BY CONDITION.....	75
APPENDIX B: IRB APPROVAL DOCUMENTS	94
APPENDIX C: SUPPLEMENTARY ANALYSES	97
REFERENCES	100

LIST OF FIGURES

Figure 1. An autonomous car approaches an unavoidable accident, and is faced with the choice between killing five pedestrians or killing two of the passengers in the vehicle. Image created by Scalable Cooperation at MIT Media Lab.....	16
Figure 2. An illustration of Fitts list. Humans have been historically better at detection, perception, judgment, induction, improvisation, and long term memory; while machines have been better at speed, power, computation, replication, simultaneous operations, and short term memory (Fitts et al., 1951).....	25
Figure 3. Estimated marginal means for moral blame for sex across agent type show an interaction wherein males, females, and other sex assign different levels of blame to different types of agents. Post hoc comparisons showed a significant difference between males' and females' assignments of blame.	49
Figure 4. Estimated marginal means for moral blame for account across apology show an interaction between account and apology, wherein coupling a neutral apology valence with a positive account valence resulted in the lowest level of moral blame.	51
Figure 5. Estimated marginal means reflecting the main effect of account on perceived trustworthiness.	54
Figure 6. Estimated marginal means reflect an interaction effect between agent type and apology on trustworthiness scores. Visual examination of the spread of scores for each agent type shows a much more consistent ranking of trustworthiness of the machine agent, whereas negative and positive valence of apology resulted in quite disparate rankings of trustworthiness toward the human agent.	55

Figure 7. Frequencies reflecting participants' perception of agency. Specifically, participants were asked if they believed the (human or machine) agent was in control of its behavior. As is reflected in this graph, people were much more likely to respond "definitely yes" for the human condition, but "probably yes" or "unsure" for the machine condition. 61

LIST OF TABLES

Table 1. List of hypotheses divided by each primary dependent variable of interest.....	37
Table 2. Hypotheses regarding moral blame and their respective outcomes.....	52
Table 3. Hypotheses regarding trustworthiness and their respective outcomes.	56
Table 4. Means and SE of participant ratings of ability, benevolence, and integrity of agents, broken down by <i>account</i> given. Note that the means presented here are based on modified population marginal mean.....	58
Table 5. Means and SE of participant ratings of ability, benevolence, and integrity of agents, broken down by <i>apology</i> given. Note that the means presented here are based on modified population marginal mean.....	58
Table 6. Coded categories of targets for displaced blame. Participants answered the “displaced blame” question organically, and the categories were created based on trends which emerged from the responses.	60
Table 7. Review of hypotheses, including whether the study presented here lends support for these hypotheses (grey boxes), or not (white boxes).	63

CHAPTER ONE: INTRODUCTION

*“I am putting myself to the fullest possible use, which is all I think that
any conscious entity can ever hope to do.”*

- HAL; 2001, A Space Odyssey (Movie)

As machines have evolved, their usefulness has also grown—allowing them to adopt meaningful roles in many areas of society. Machine agents (that is, machines with some form of nascent “intelligence”) have been developed to aid in domains such as military command and control by collecting and sharing information to facilitate strategic military missions (McGrath, Chacón, & Whitebread, 2000). Such agents have also been developed to help people in many differing contexts. For example, a robotic agent has been developed to help blind people navigate busy streets (Harris, 2015); a multitude of intelligent personal assistants have been designed to help organize and find information for people (e.g. Apple’s Siri, Google’s “Google Now,” and Microsoft’s Cortana). As a result of these endeavors, humans are not only relying more on machine agents, but also beginning to see them as potential teammates and friends (Ososky, Schuster, Phillips, & Jentsch, 2013; Piore, 2014). Many people in modern society eagerly celebrate this dawn of the “Fourth Industrial Revolution”—which is driven by the universality and versatility of intelligence that exists in technology today (Schwab, 2016).

The adoption of machines into these meaningful roles has resulted in a number of controversies—particularly regarding their potential for harm. Many luminaries and skeptics have voiced concern about the risks and dangers of letting machines have untrammelled intelligence, lest they use that power to make humans obsolete (Bostrom, 2014; Future of Life

Institute, 2015; Hancock, 2017; Stowers, Leyva, Hancock, & Hancock, 2016). Indeed, there is already the potential for harm in machines with no intelligence at all. For example, while drones can be used in search and rescue operations (Muoio, 2016), they can also be used to make strikes on enemies in warfare (Woods, 2015). Even machines such as surgical robots, designed to help people, have accidentally killed patients whilst being used (BBC, 2015). But machines have always had an element of risk inherent in their physicality. Adding intelligence to machines introduces greater complexity, and the potential for greater risks. For example, in May, 2016, a Tesla vehicle with a relatively simplified level of “intelligence” (i.e. it could sense its surroundings and act on them) struck a tractor trailer while essentially driving itself (using autopilot; see Muoio, 2016). While the vehicle was supposed to be under supervision by its driver, the accident occurred with no interference from the driver, leaving society to question the locus of fault.

Such harm as that detailed above has not prevented society as a whole from using machines, or even relying on them, on a daily basis. Indeed, some might argue that modern civilized society has little choice. Citizens continue to use, rely on, and relate to machines of varying levels of intelligence with seemingly little regard for potential disasters, at least until after they come to light. At this point, judgment is passed on the perpetrator—be it machine or human—that is seemingly responsible for the disaster. Complaints of people relying on machines *too much* only arise when disaster strikes (Carr, 2013). For example, immediately following the Tesla accident mentioned above, news media sparked a heated discussion surrounding an unverified claim that the driver of the Tesla vehicle was watching a movie instead of supervising his vehicle (Levin & Woolf, 2016).

Herein lies a gap in our understanding of human attitudes toward machines. What is the nature of human judgment toward machines when they *are* responsible for disaster, especially deadly ones? More importantly, to what extent are these machines *blamed* for their behaviors? As intelligent machines are now an integral part of society's triumphs and tragedies, it is necessary to examine blame in this context to understand and increase their safe acceptance in society. After all, the process of forming a judgment or evaluation of a person lays the groundwork for establishing trust in human relationships (Mayer, Davis, & Schoorman, 1995; Schoorman, Mayer, & Davis, 2007). Likewise, breakdowns in trust and relationships are embedded in ongoing judgements—including the placement of blame—regarding behaviors (Kim, Dirks, Cooper, & Ferrin, 2006). Such effects may also arise in human interaction with machine agents, especially when lives are at stake and morality becomes a driving design factor.

At present, our understanding of moral blame toward machines remains limited. We have not specified how such blame emerges and how it can aid or impair acceptance of machines. However, human judgment and blame of other humans may offer a parallel for understanding how they may blame machines for morally-laden decisions. Much research has been completed on the concept of *moral judgment*—herein defined as the evaluation of “right versus wrong”—and its relationship to *moral blame*—or the perceived blameworthiness of an individual for engaging in a behavior that is judged to be “wrong” (see Pizarro, Uhlmann, & Salovey, 2003). This research, often conducted using written vignettes describing various morally conflicting behaviors which participants must react to, has helped in the understanding of human placement of blame and how it can be mitigated or exacerbated. Yet, the findings to date present a rather complex picture of human blame. For example, research has shown that even harmless actions

can be judged as morally blameworthy if the actor in question benefits from misfortune caused by the action (Inbar, Pizarro, & Cushman, 2012). Research has also shown that if such an actor is portrayed as having done something wrong intentionally (as opposed to accidentally), perceived blameworthiness is not only higher, but the perception of the crime itself is also greater in magnitude (Ohtsubo, 2007; Pizarro, Laney, Morris, & Loftus, 2006).

It is possible that such research can be applied to our understanding of human judgments of machine behavior, since some work has already shown potential parallels between the two. For example, social attraction theories governing the response of humans' acceptance of each other also govern humans' acceptance of machine agents (Nass, Moon, Fogg, Reeves, & Dryer, 1995). However, humans don't always respond to machines in the same way they respond to each other. They don't even respond to less humanlike machines the same way they respond to more humanlike machines, as has been found in research showing that contrasting trust constructs differed according to the "humanness" of machines being interacted with (Lankton, McKnight, & Tripp, 2015). Effects such as these reinforce the caution by some researchers against the creation of machines that are too humanlike, lest such machines instill undeserved expectations from human users and teammates, who may appraise them as being capable of a higher standard of behavior than they actually are (Norman, 1994). After all, humans' appraisals of each other are markedly different from their appraisals of machines, leading to quite different expectations. For example, who would expect a modern-day machine to love or understand compassion? It is possible this same difference in appraisal and expectation toward humans and machines exists in moral blame, which is inherently based in human expectation of others' behaviors. Herein lies the primary question this dissertation will examine.

Research Purpose

Given the foregoing questions and gaps noted here, the purpose of this dissertation is to explore differences in moral blame attributed to human and machine agents. Specifically, this dissertation will examine whether the knowledge and type of *reason*, as well as *apology*, for an untoward incident mitigates moral blame against the parties involved. Furthermore, the effect of such blame on trustworthiness will be evaluated. In order to fill the gaps in understanding between topics in moral psychology, cognitive psychology, and artificial intelligence, valuable information from each of these fields are being combined to guide the research studies being presented herein.

CHAPTER TWO: REVIEW OF LITERATURE

“Morality is not just any old topic in psychology but close to our conception of the meaning of life. Moral goodness is what gives each of us the sense that we are worthy human beings.”

- Steven Pinker

One of my key goals here is to identify differences in how humans react to human agents versus machine agents when faced with accidents caused by these agents. To address this, several topics are examined in detail. First, a conceptualization of the term “agent” is presented in order to differentiate between current human and machine agents. Next, topics concerning morality, including nascent concepts in machine morality, are considered. This leads to a discussion on human attributions of blame toward other humans and nonhuman entities. In order to completely understand human blame in this context, it is necessary to consider the process of human moral judgment. Additionally, the role of professed reasoning and apologies in mitigating emotional responses and attributions of blame toward agent behavior is explored. These discussions culminate in a series of research questions and overall research objective—which is addressed through the implementation of an in-depth research study.

Agent to Agent

Before making direct comparisons between human reactions to human agents versus machine agents, it is prudent to define what an *agent* is, what it is not, and how the term is used for the present purpose. The term *agent* originated in the Latin word *agere*, meaning *to do* (Agent, 2017). The use of the term has evolved over the centuries to represent several different concepts, including “one who acts,” “any natural force which produces a phenomenon,” as well

as a “representative”. Since the early 1900s, this term has adopted further and more varied meanings specific to various roles humans may play. For example, today it is common to hear of all types of agents, including “secret agents” and “insurance agents.” It has also become common to see *agent* representing nonhuman entities (e.g. “chemical agent”). When considering the evolution of the term, it is not at all surprising that *agent* now includes several types of machines.

While all humans can be considered agents, not all machines are agents. The specific designation of agent most commonly applies to machines that are intelligent enough to *perceive* and *act on* their surroundings with some degree of autonomy or independence (Russell & Norvig, 2009). This definition aligns well with the Latin origin of agent as it includes the requirement of *action*. It can also be applied to both humans and machines. Using this definition, Russell and Norvig (2009) make a direct comparison between human agents and machine agents: while a human agent perceives through eyes, and ears, etc., machine agents may perceive through cameras and other electronic sensors. Similarly, while human agents act on their surroundings using arms and legs, machine agents use motors, gears, algorithms, and other effectors. In this way, both human and machine agents exert action in the world.

Many types of machines may be designated agents according to Russell and Norvig’s (2009) definition. For example, robots such as the Nao robot—which can see, hear, feel, speak, move, “think,” and even exhibit a semblance of self-awareness—would fall under this definition of agents (Soft Bank Robotics, 2017; Pandey, 2015). Software programs such as Watson (IBM, 2016) and Deep Q-Network (Mnih et al., 2014; Pandey, 2015) can also be considered agents as a result of their capabilities in sensing, problem solving, and communicating with humans. Self-

driving cars have also begun to shift into the roles of “agents” as they have started making decisions based on their perceived surroundings. Machines that would not be considered agents under this definition include teleoperated robots such as those used for explosive ordnance disposal (EOD) and surgery (e.g. da Vinci; see Siciliano & Khatib, 2008). While robots such as these have the capability to *sense* their surroundings, they cannot currently *act on* their surroundings autonomously. Even though their sensing capabilities afford an appearance of intelligence beyond other machines, the fact that they merely act as an extension of human movement prohibits their inclusion into what Russell and Norvig would call “successful agents” (Russell & Norvig, 2003, p. 32).

Within the present scope, the term agent is used to signify either a human (specified “human agent”) or a machine that *senses* and *acts on* its surroundings (specified “machine agent”; Russell & Norvig, 2003). Teleoperated robots are not included in this definition. Defining agents in this way maintains a focus on human interaction with cutting-edge machines that are so intelligent, they possess enough autonomy to take responsibility for their actions, or at least be blamed for them.

Agent Autonomy

What is “autonomy” and how much constitutes “enough to take responsibility”? Classically defined, autonomy (origin Greek, “autonomos”) is the freedom of control or government of oneself (Autonomy [Def. 1], n.d.). Inherent in autonomy is the notion of moral independence, or the ability to act rightly or wrongly according to one’s own desires (Autonomy [Def. 2], n.d.). As a price for freedom and self-government, autonomous beings are granted

responsibility—and often (but not ubiquitously) held accountable for their behaviors. However, just as in the case of “agents,” modern-day definitions of autonomy have evolved to encompass entities that didn’t exist at its linguistic conception—in this case, machines. However, autonomy in humans is typically discussed in a manner entirely different from that in machine agents. Whereas autonomy in humans is frequently viewed as a developmental characteristic that changes naturally across the lifespan, autonomy in machines is both planned and designed, typically with some forethought concerning how the machine’s autonomy relates to its human counterparts.

For example, in humans, one might consider that one has less autonomy in childhood than in adulthood. Further, some parents allow their children more autonomy as they age (i.e., promoting choice) while other parents are more controlling (i.e. pressuring them toward certain outcomes; Deci & Ryan, 1987). Regardless of the amount of autonomy any human is allowed, it is typically assumed that most humans can—and often do—attempt to achieve higher levels of autonomy as they mature. For example, a parent may allow one child to have more autonomy than another, but either child could attempt to exert a higher level of autonomy than she currently possesses. Indeed, humans do this throughout childhood, particularly in the first few years when they learn to walk and explore their surroundings (Erikson, 1963; Stapel-Wax, 2011). Even animals, which may appear to have less autonomy than humans, act in this way.

On the other hand, autonomy in machines is defined and implemented in a stricter sense. Autonomy in machines requires the capability to independently compose and select actions to accomplish goals based on an understanding of the world, including a machine’s own role in

specific situations (Parasuraman, Sheridan, & Wickens, 2000; Shattuck, 2015). Since autonomy is designed and created in machines, they typically express greater limits in the extent to which they can vary their own behavior. Yet, autonomy in machines continues to evolve over time. This evolution is closely linked to success in Artificial Intelligence (AI), the capability of machines to perform tasks requiring increasing amounts of intelligence (DoD, 2016). As successes in AI allow machines to approach human-like autonomy, new problems are arising, including those explored herein. This work evaluates machines at the cutting edge of this evolution, specifically examining human reactions to machines which at first glance appear to be just as autonomous as adult humans. It is in this one-to-one comparison that human reactions to machine morality and potential blame will be most evident.

Whose Morals?

It is because of humans' ability to act autonomously that we adhere to notions of morality in our lives, including the assessment of morality in others. But what is morality? The term *morality* is born of the Latin word *moralis*, or the proper behavior of a person in society (Morality, 2017). Morality generally involves the distinction between "good" and "bad," and can thus be thought of as rules that govern a society. In this way, morality acts as the spine or fiber that holds a society together. This can be demonstrated in the idea of *tribalism*, or the tendency for humans to favor, and often protect, those in their immediate social circle (Greene, 2014). It can be further distinguished through humans' general unwillingness to harm even those they *don't* know (Cushman, Gray, Gaffey, & Mendes, 2012).

Until recently, researchers believed morality was a trait specific to humans. Even Darwin, who was well ahead of his time in his understanding of human and animal life, suggested that morality would only be found in beings with intellectual capabilities approaching that of humans (Darwin, 1888). However, it has been suggested that other primates may possess morality, albeit in a limited way, which they exhibit through compassion and concern. For example, in a controlled laboratory experiment, Warneken and colleagues found that chimpanzees were willing to help each other, as well as humans, without expecting rewards (Warneken & Hare, 2007; Warneken & Tomasello, 2006; Warneken & Tomasello, 2009). Examples such as this abound in research on nonhuman primates (Greene, 2014). Similar examples can be found in other animals as well, including carnivores (e.g. wolves), cetaceans (e.g. dolphins), and, even some rodents (e.g. rats; Bekoff & Pierce, 2009). Beyond the possibility of naturally-existing morality, there have also been discussions regarding the opportunity to create morality in human-nonhuman chimeras, which has sparked much ethical debate (Piotrowska, 2014). Still more difficult and controversial is the possibility of creating morality, or at least a semblance of it, in machines (Hancock, 2009; Stowers, Leyva, Hancock, & Hancock, 2016). In many ways, this can be thought of as the modern frontier of morality, especially if one considers all products of creation to inherently possess a moral dimension (Hancock, 2009). But to even broach this frontier, it is first important to consider how morality has come to exist in humans, and how it can possibly begin to exist in machines.

Morality in Humans

The origin of human morality is difficult to pinpoint since it occurred before the beginning of recorded history. However, theories abound concerning not only its origin in

humans as a species, but its development in individuals. Human morality is often presumed to have evolved in much the same way as other survival traits—as a characteristic that aided in “survival of the fittest”. Some scholars argue that genes which encouraged moral behavior became more prominent in humans because moral humans developed more effective social support systems which protected them (Broom, 2003). Other scholars suggest that the evolution of morality in humans grew largely as a result of brutal social control that may have emerged 45,000 years ago (Boehm, 2012). The stricter the enforcement of this control (which often included capital punishment), the more likely was the social selection of morality (i.e. behaving according to expected social norms). Those who didn’t act according to social norms were “eliminated,” while the rest survived and passed on genetic and behavioral traits necessary to successfully survive in tribes. It is difficult to test either of these hypotheses. However, research on Late-Pleistocene Appropriate (LPA) hunter-gatherer societies that exist today gives some credibility to the latter (Boehm, 2012). This idea of morality-by-sanction is also consistent with modern-day use of punishment for controlling inappropriate behavior, making it a promising explanation for modern human morality, and perhaps even prospective machine morality.

As difficult as it has been to reach consensus on the moral evolution of the human species, the moral development of individual present-day humans—while much easier to study—has proven equally as difficult to understand. Kohlberg (1973) postulated that moral reasoning, specifically defined as one’s sense of justice, develops in 6 discrete stages that last throughout one’s life (Kohlberg, 1973). The linear development of moral reasoning according to this theory provides support for the idea that as children gain moral competence (i.e. understanding right versus wrong in increasingly complex situations), they are really gaining moral autonomy. Such

a notion aligns well with our understanding of the allowance of behavioral autonomy in children discussed above. However, conflicting evidence arose regarding Kohlberg's original proposed developmental stages, causing other researchers to begin operationalizing morality and moral development in increasingly varied ways. Turiel and colleagues (1983) proposed the Domain Theory, which explains human moral development in connection with societal and psychological development. According to this theory, morality is also operationalized in a much broader sense than simply justice, and includes concepts such as fairness and equality (Turiel, 1983). More recent work has broadened the scope of morality even further, as can be seen in Moral Foundations Theory, which explicitly seeks to define and measure morality across five foundational dimensions: care, fairness, loyalty, respect, and sanctity (Graham et al., 2011, 2012). Moral Foundations Theory, originally developed to understand morality across cultures, has successfully highlighted that morality is much more complex than even envisioned by Turiel.

Today, we know that the manifestations of morality vary across history, religions and context. Local and religious morality systems may dictate different rules on the respect of deities and cultural customs (Greene, 2014). Within a single system, morality may still change over time. For example, attitudes on the morality of various sexual tendencies and behaviors in Western societies changed substantially in the second half of the twentieth century relative to the first half (Stewart-Williams, 2010). This can be evidenced in the de-criminalization and wider acceptance of homosexuality, as well as a wider acceptance of promiscuity. This ongoing variation in morality makes it difficult to study, and even more difficult to understand. In such an ever-changing and diverse world, is there any such thing as fixed moral precepts? Where will morality take us next?

Morality in Machines

The coming frontier in morality may very well look to apply morality to machines. But can machines possess the same morals humans do? At present, no. While machines are programmed to follow a set of preset guidelines, these guidelines do not operate the way human morals do. They're simply a set of directives, instructions, or algorithms, leaving the machine computationally restricted in their behaviors (Brundage, 2014; Stowers, Leyva, Hancock & Hancock, 2016). Furthermore, machines do not exhibit emotional conflicts when they have to take a life. Nor do they necessarily weigh the cost of life the same way humans do (i.e. with an emotional element). While attempts have been made to design "moral emotions" into machines (e.g. guilt), such emotions are algorithmically based and not organically developed or learned (see Arkin, 2011; Arkin & Ulam, 2009).

The idea of implementing morality in machines is not a new one. This topic has gained increasing attention over the years (Hancock, 2009; Hancock, 2017), especially as machines have grown enough in intelligence to make decisions that shape human lives. Take, for example, the growth of automation in cars. What began as anti-lock braking systems to save humans from needing to repeatedly apply the brakes in quick succession, and electronic stability control to prevent humans from loss of control on slick roads, has now progressed to autopilot functions which steer, merge, and brake for humans. Fully automated commercial road vehicles are also now being tested, with their widespread implementation coming in the near future (Sage & Lienert, 2016).

The introduction of fully automated vehicles to mainstream public traffic has led to a new series of questions, not altogether unlike those already studied in human morality. For example, after it became apparent that autonomous cars are a realistic part of our future, several information sources called for research examining how cars should react in “trolley problems” (Lin, 2013). In the traditional trolley problem, a human is faced with an unavoidable collision and must decide whether to let the trolley kill five people on the tracks it is currently set on, or switch it to kill just one person on another set of track. In the machine version of this problem, an autonomous car will be faced with a choice between two collisions (avoiding one will cause the other), and will have to decide how to respond. As in the traditional problem, someone will die no matter what happens. Figure 1 illustrates an example of such a collision where there is no escape. Should the car save its passenger no matter what, or should it instead prioritize the number of lives that will be saved overall?

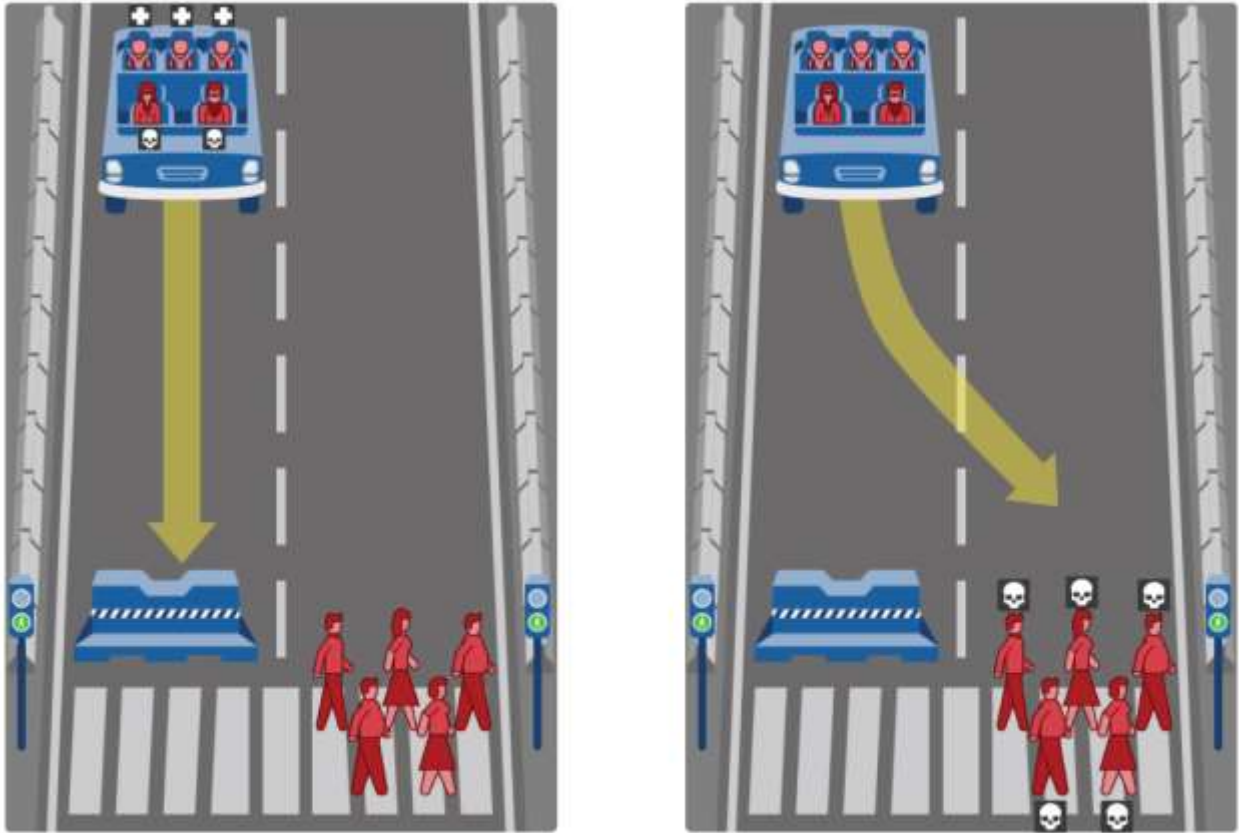


Figure 1. An autonomous car approaches an unavoidable accident, and is faced with the choice between killing five pedestrians or killing two of the passengers in the vehicle. Image created by Scalable Cooperation at MIT Media Lab.

A recent series of studies examined human reactions to this dilemma and found that we as humans don't have a consistent opinion about how this should be resolved (Bonnenfon, Shariff, & Rahwan, 2016). Results from these studies showed that, while participants believed that autonomous vehicles should prioritize the number of lives saved over their own passengers, they would not be comfortable being in a vehicle which operated under this rubric. Instead, they preferred to be in vehicles which always prioritized their own passengers' lives over the lives of others. Unsurprisingly, given the conflicting feelings humans have about how these dilemmas should be handled, participants also stated that they didn't want responses to such dilemmas by the machine to be mandated legally.

As enticing as the trolley problem sounds, the reality is machines aren't designed to deal with such moral dilemmas because they are unlikely to arise in a machine-driven world (Hern, 2016). The design of machines for everyday situations such as driving or handling office work is typically centered on creating a level of perfection that won't allow for the mistakes that lead to killing someone. On the rare occasion that such a dilemma does arise, the machine's programming is most likely to have a directive to stop rather than swerve (if a self-driving vehicle) or behave in as conservative a way as possible (if in a healthcare or workplace setting).

Yet, even though machines likely won't have to deal with "trolley problems," they will be faced with other types of moral dilemmas. For example, some researchers have highlighted the concern that, even though programmers will always try to account for as many situations as possible when creating machine agents, there will always be a point at which a machine agent will be faced with a situation it hasn't encountered, and this point may lead to harm for humans involved (Scheutz, 2016). This has already happened, as witnessed in the case of the Tesla automobile accident in Florida (Muoio, 2016). Such adverse events will likely continue to happen across many contexts, including worker and domestic life. This realization has led many to encourage the creation of "implicit ethical agents," i.e. those with ethical and/or moral guidelines for moral emotions (e.g. guilt and compassion) implicitly built into their design (see Arkin, 2008, 2010; Moor, 2006; Scheutz, 2016; Stowers, Leyva, Hancock, & Hancock, 2016). An extension of such a goal would include the somewhat organic evolution of these guidelines through sophisticated machine learning methods. After all, humans already appear to be holding machines to higher standards of moral and ethical behavior, and this trend will likely continue.

Playing God

Humans' ability—and even desire—to behave “morally” lends them the ability to pass judgment on the morality of others' behaviors and assign blame where they believe others should take responsibility. Just as humans have opinions about how one should respond to a trolley problem, they also have opinions about mistreating others, causing pain, or killing—whether intentionally or unintentionally. After all, this is fundamentally what the area of jurisprudence is about. Yet, even these opinions become very inexact and confusing when examining their role in blame toward those who exhibit morally questionable behavior. This phenomenon, the accusation that someone is causally responsible for seemingly immoral behavior, is often given the label “moral blame.” But how does moral blame work, and what are its underpinnings?

Moral blame, first and foremost, exists as an extension of moral judgment—judgments of how people should treat each other (Turiel, 1983). Such judgments can be rendered in response to trolley problems or other dilemmas where the perpetrator may be more directly responsible (Greene, 2001). Most importantly, these judgments are passed on everyday behaviors—toward a manager dealing with clients, toward a doctor dealing with patients, and toward other drivers on the roadway. Researchers have examined moral judgment in detail in order to understand how it happens and how it is then linked to moral blame.

One of the hotly debated questions concerning moral judgment is whether the process of passing initial judgment (e.g. is that moral or immoral?) is intuitive (automatic-affective) or rational (reasoning-based). While many classic theories of moral development focus on the role of reason (e.g. Kohlberg, 1973), such research when compared to more automatic affective

reactions has historically yielded mixed results (e.g. Galotti, 1989). After decades of confusion surrounding this conflict, the modern consensus is that moral reasoning actually requires both. Greene and Haidt (2002) posit that people may have an automatic reaction in judging events as moral or immoral, but they justify their automatic reactions through post hoc rational support. They may be motivated to do so in order to defend against threatening ideas (Greene & Haidt, 2002). These conclusions are backed by both behavioral and neurobiological data (Adler & Rips, 2008; Greene, 2001). Yet our understanding of exactly how these processes occur and evolve remains primitive at best. Given the complexities involved in moral judgment, it is even more difficult to understand its relationship to moral blame.

Pizarro and colleagues have examined moral blame directly through several studies in which participants are presented with a narrative explaining that someone has engaged in some type of immoral behavior (e.g. intentionally smashing someone's window, walking out of a restaurant without paying; Pizarro et al., 2006; Pizarro et al., 2003; Pizarro, Uhlmann, & Bloom, 2003). Findings from these studies have highlighted several interesting trends in the way people assign moral blame. Moral blame is discounted in situations when the negative behavior appears to be impulsive rather than deliberate (Pizarro et al., 2003). Furthermore, moral blame toward others is discounted when intentions and outcomes don't necessarily match (e.g. intending to kill someone, but only accidentally killing them; Pizarro et al., 2003). Most notably, moral blame in such situations is more likely to be attenuated when assessed intuitively, rather than rationally. Not only does this highlight the complexity of the intuitive versus rational moral judgment debate, it also creates the possibility that moral blame can be moderated by rationalizations of behavior. Such a possibility is one that is considered as part of this dissertation. But first, it is

prudent to consider whether the findings concerning blame toward human agents hold true for machine agents also—as it is in this primary comparison of human and machine agents that the purpose of the present work lies.

Equal Blame

Do humans judge machine agents the same way they judge each other? Should they? Given the many differences between humans and machines (e.g. humans are biological and machines are typically not, humans are conceived as possessing “free will” and machines generally are not), it seems logical that blame toward the two could not be identical. But what about in the case of very intelligent machines—particularly machines argued to be as intelligent or nearly as intelligent as humans? Or in the comparable case of humans deemed lacking in enough intelligence and maturity to receive diminished responsibility? In order to explore answers to such questions, it is important to first consider the following question: Are machines as blameworthy as humans?

The first requirement for someone to be responsible or blameworthy for a behavior is that they have control over the behavior (Tognazzini & Coates, 2014). As such, many morally questionable behaviors that are completed under circumstances where the perpetrator has either limited, compromised or complete lack of control over the behavior will often be judged less harshly (e.g. in “crimes of passion” such as violence upon witnessing adultery; Pizarro et al., 2003). When considering how blameworthy any machine is, particularly in comparison to a human, it is necessary to consider whether the machine has equal control to a human over the behavior. While machines don’t exhibit the sort of lack of self-control that results in emotional

“crimes of passion,” it can still be argued that they have less control over themselves than humans do. Indeed, any control machines have over themselves is primarily directed by humans through software programming—telling them what they are “allowed” (or even able) to do in certain circumstances.

A useful, albeit imperfect, analog to the concept of self-control in machines is self-control in children. As discussed earlier, children possess less autonomy than adults and may be allowed more or less autonomy by their adult guardians. Furthermore, this allowance of autonomy often aligns with moral development as defined by Kohlberg (Kohlberg & Hersh, 1977) and Turiel (1983). The consequence of this social “allowance” of autonomy for children is that adults are frequently deemed more morally blameworthy than children when untoward events occur. Indeed, parents will blame each other (and often themselves) for children’s behavior, often operating on the sense that children’s behaviors are reflections of successful (or failed) parenting (Drexler, 2012). Companies have even capitalized on this tendency by assigning blame to parents for things that are really the fault of the companies themselves (e.g. when public health officials in the U.S. tried to enact legislation preventing the addition of lead to paint in order to prevent brain damage in children, the lead industry blamed uneducated parents for letting their children put lead-contaminated items in their mouths; Rosner & Markowitz, 2013). The general assumption is that children are not entirely responsible for their behaviors, and thus they are not assigned all of the blame for their behaviors; some (or all) of that blame is placed on the guardians instead.

In many societies, this attribution of less blame toward children is upheld legally through the assignment of milder consequences. Take, for example, the existence of juvenile court in the United States. A citizen under the legal age of adulthood (a “minor”) who commits a crime is typically tried in juvenile court—often with milder consequences if found guilty (Scalia, 1997). Some particularly gruesome crimes committed by minors, especially as those minors age and develop, may still result in them being tried in traditional courts. Yet, even under such circumstances, the decided punishment, especially if it involves life in prison or death, is considered highly controversial (Kirkland, 2012). Furthermore, some states in the US (alongside some countries in Europe) will still authorize punishments to parents of offending minors, on the basis that they were legally responsible for the minors at the time of their crimes (Le Sage & De Ruyter, 2008). This suggests that, in humans, the alleged blameworthiness and assignment of blame toward children (i.e. humans with less presumed autonomy) and adults (humans with more presumed autonomy) can be quite fluid.

Autonomy in animals is far more straightforward than it is in humans, and may be a more appropriate analog to understanding self-control in machines. Thus it is necessary to consider how societal views of animal behavioral issues may apply to such views of machine behavioral issues. Societal expectations of animals are altogether different from their expectations of each other. Whereas autonomy in humans is largely linked to the developmental stage in their lifespan in relation to other humans, autonomy in animals is constrained by not only their development, but their intelligence and their domestication. For examples, animals that are not domesticated could be argued to have more autonomy than those who are. Similarly, a highly intelligent dog

could be argued to have more autonomy than a snake—simply because it has a better understanding of its surroundings and how to enact its own force on the world.

Many have argued that animals have enough autonomy as beings that they should have *rights* in society (Sunstein, 2003). Yet, even while arguing for the rights of animals, not all in society agree that they should be given *responsibility* in equal measure. For example, if a dog attacks a child, many argue that the blame for the attack lies with the dog’s owner for mistreating the dog in ways that may lead it to attack. This is a perspective supported by organizations such as People for the Ethical Treatment of Animals (PETA), which exist to protect animal welfare (e.g. see PETA, 2017). Others argue that both the dog and the owner involved in such an attack should be punished accordingly—the dog through death, the human through a fine or incarceration. Laws vary in how they enforce these differing societal views, with some cities and states have specific statutes in place for handling dog attacks (Snyder, 2017). However, one thing remains consistent: much like the case of children, bad behavior by domesticated animals generally results in either alleviations of blame altogether or otherwise supplemental blame (i.e. blame toward someone other than the perpetrator). Thus, the question arises: where do machines, their creators, and even their owners fall as parties deserving of blame when a machine makes a mistake? Are machines “pets” and their creators or bosses “owners”? Are designers “guardians”? Unfortunately, there is no simple answer.

Machines don’t have the same level of autonomy as adult humans. Much like children and animals, they are also not accorded full responsibility for their actions by society. Take, for example, the aforementioned Tesla accident. An investigation by NHTSA sought to pinpoint any

cause for blame toward either the automobile or the manufacturer for the accident (NHTSA, 2017). At the close of their investigation, NHTSA reported that the vehicle, despite failing to see the truck it crashed into, did nothing wrong and had operated within the full limits of its capabilities. NHTSA also found no cause for blame toward Tesla, though many citizens would argue that Tesla, as the “parent,” should take responsibility to ensure drivers properly use their vehicles (Ohnsman, 2017). Indeed, given what we are learning about the ability of humans to effectively monitor semi-autonomous driving systems (Endsley, 2017), companies such as Tesla would do well to recognize humans’ limitations and accommodate them accordingly, rather than pinning the blame exclusively on the humans themselves.

From a societal and legal perspective, lack of blame toward machines isn’t unexpected. After all, computation in the machines we know today has only existed in recent history, with the first sign of modernly-defined artificial intelligence being created in the 1940s (Russel & Norvig, 2009). Since then, machines have gained intelligence rapidly, yet have always fallen short of human intelligence in the most complex areas of intelligence. For example, consider the list, developed by Paul Fitts and colleagues (Fitts et al., 1951), detailing the activities humans are better at versus those machines are better at (see figure 2). While Fitts asserted that machines are better at activities which require speed, power, computation, replication, multi-tasking, and short-term memory; humans are better at detection, perception, judgment, induction, improvisation, and long term memory. It has since been suggested that machines now also surpass humans in detection, perception, and long-term memory (Bostrom, 2014; de Winter & Hancock, 2015). This gives credence to the notion that machines should certainly be held to a higher standard now than they have been historically. Yet it doesn’t address other areas of human

intelligence that weren't necessarily accounted for in Fitts list—areas that are much more relevant today now that machines are being trusted to make decisions. One such area is moral and ethical intelligence.

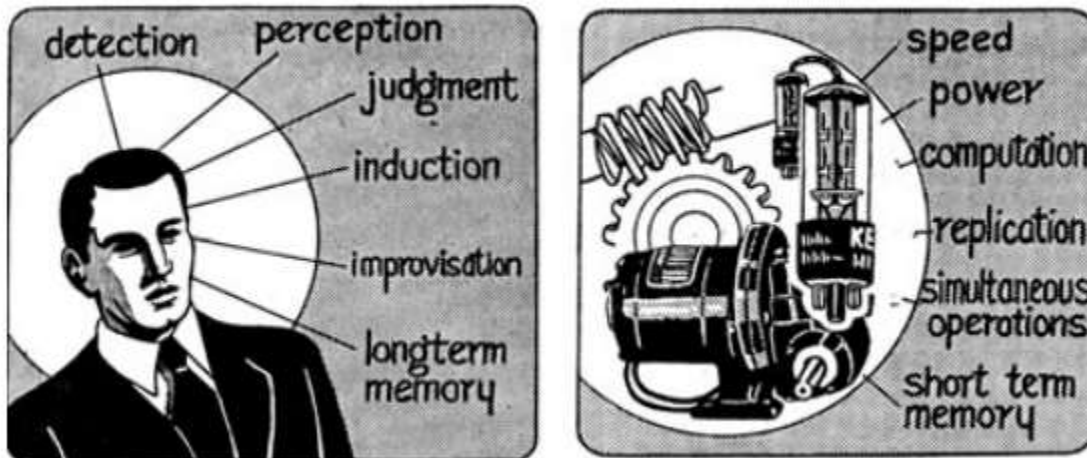


Figure 2. An illustration of Fitts list. Humans have been historically better at detection, perception, judgment, induction, improvisation, and long term memory; while machines have been better at speed, power, computation, replication, simultaneous operations, and short term memory (Fitts et al., 1951).

Scheutz suggests that it is through the design of machines as “ethical agents” that humans can start judging their moral competence (Scheutz, 2016). It is also through such design that the answer of whether machines should be blamed for their mistakes—especially morally impactful ones—could be made much more straightforward. Dennet (1997) proposes that the primary characteristic missing for machines to be culpable for any wrongdoing is *mens rea*, or knowledge of wrongdoing when engaging in the behavior. *Mens rea* is the last frontier in machines that must be crossed for machines to be held truly accountable for their actions in society. What remains unclear is how this accountability will be upheld. Will machine agents ever be “punished” for wrongdoing? What would such punishment look like?

Machines may not yet have *mens rea*, but they do possess increasing amounts of autonomy, and it is this autonomy that opens them up to blame in everyday interactions, regardless of their legal culpability. Furthermore, the average citizen interacting with machines isn't necessarily aware of a machine agent's lack of *mens rea*. Indeed, many people believe machines possess intentionality (Friedman & Millet, 1995). Research has also shown that human belief in machine capabilities may extend beyond *mens rea* to other qualities inherent to human agents. The idea that humans may apply social rules to machines is known as "Computers Are Social Actors" (CASA; Nass et al., 1995), or the Theory of Social Responses to Computers (Reeves & Nass, 1996). This theory and related research contends that humans' interactions with machines are inherently social. As a result of this social interaction, humans categorize machines as social actors much like themselves and may judge them with the same favor (Broadbent, 2017; Nass & Moon, 2000).

This theory of machines as social actors may explain some trends in task-based blame toward machines that already exist. For example, Hinds and colleagues (2004) found that, in interacting with robots on a joint task, people are more likely to blame robots for low team performance when the robots appear to be in positions of elevated status (i.e., supervisors; Hinds, Roberts, & Jones, 2004). It was similarly found that attribution of blame is heightened toward robots that are more autonomous (Kim & Hinds, 2006). This suggests that, as machine agents continue to gain reputations for having intelligence, increasing levels of blame will be placed on them. These findings have also been upheld in research examining moral blame in morally charged decisions. Research comparing moral blame toward machine and human agents has found that autonomous machines with the capacity to make decisions are certainly targets for

moral blame in moral dilemmas (Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). Machines overall received the same amount of blame as humans for their responses to moral dilemmas. The primary difference between them was that machine agents typically received more blame for *inaction*, while humans received more blame for *action*. The narrative examined in this dissertation involves a decision that results in *action*.

Here, I examine whether judging machines as social actors extends to machine agents in control of self-driving cars—a technological area fraught with controversy in society. Whereas a theory of social actors might support that there should be no difference in moral blame toward human and machine agents, my first hypothesis explores whether the alternative could be true. Given that the scenario included herein involves a morally-charged mistake resulting from *action*, it is possible that human perception of that mistake may vary when completed by a human agent compared to a machine agent. Furthermore, given that society has already begun to form opinions on the culpability of (or lack thereof) self-driving cars involved in accidents, it is worth considering how their views translate in such a scenario as the one examined here. Thus, my first hypothesis indicates that a morally-charged mistake made by a human agent will result in a higher attribution of moral blame compared to a machine agent making the same mistake (see Table 1).

Should I Trust You?

A primary impact of moral judgment and moral blame lies in the establishment of trust toward agents being judged or blamed. The assessment of an agent's moral character is partly an assessment of one's *trustworthiness* to behave morally. After all, trustworthiness is considered

one of many characteristics that encompass one's moral character (Pizarro & Tannenbaum, 2011; Goodwin, Piazza, & Rozin, 2014). But what is trustworthiness and how are perceptions of trustworthiness impacted by moral blame?

Trustworthiness, as defined by Mayer and colleagues (1995), is the complex interplay of three factors in an agent that will lead it to be more or less trusted. Those three factors have been primarily identified as 1) ability, 2) benevolence, and 3) integrity. Ability refers to the skills and competencies that allow an agent to be successful in a domain, and involves little to no assumptions on the morality of that agent. Benevolence and integrity, however, are largely moral in nature. Benevolence—or the degree to which an agent wants to do good—and integrity—or the set of acceptable principles an agent adheres to—are largely centered on ideas of right and wrong (i.e., morality). While each of these three factors may vary independently, they form a complex interrelationship which is the basis of perceived trustworthiness in agents (Mayer et al., 1995).

Judgments about trustworthiness in others are usually made quickly, much more quickly than characteristics such as likeability. Thus, trust judgments are a part of the *first impressions* humans make of others when meeting them (Willis & Todorov, 2006). From an evolutionary standpoint, such swift judgments of trustworthiness are necessary, as it helps humans to avoid agents who may harm (Pizarro & Tannenbaum, 2011). Perceptions of trustworthiness can also be influenced by other factors; particularly in relation to morality. For example, in a series of studies examining perceptions of human agents' moral judgments, participants rated agents as more trustworthy when the agents' judgments were more deontologically driven (i.e., driven by

moral duty; Everett, Pizarro, & Crockett, 2016). This suggests that humans not only make judgments regarding others' morality; they base their perceptions of trustworthiness on it as well.

Research regarding perceptions of trust in and trustworthiness of machine agents compared to human agents is somewhat mixed. While some question whether trustworthiness in its classical form (e.g., ability, benevolence, integrity) is ever perceived of machines (Friedman et al., 2000), a great deal of research suggests otherwise (e.g., Cassell & Bickmore, 2000; Komiack, 2003). Specifically, it has been found that trust and trustworthiness is quite similar toward human and machine agents, with differences lying mainly in precisely *how* it is developed (Benbasat & Wang, 2005; Komiak, Wang, & Benbasat, 2004). This research, alongside the CASA theory (Nass, et al., 1995), lends support to the idea that perceived trustworthiness of human and machine agents shouldn't be that different. However, the studies that found these similarities focused largely on computer interfaces and web-based agents. Researchers have yet to say how these findings transfer to other contexts.

Once again, it is important to consider whether a machine agent driving a vehicle will be considered in the same light as a human agent driving a vehicle. Can a self-driving car be trusted, or is it trustworthy, if all it does is drive? Do attributes such as benevolence and integrity apply to self-driving cars? If humans attribute a wider range of moral qualities to other humans than machines, then it may follow that they will be more likely to find other humans trustworthy, especially in the case of self-driving cars, which society may view as having less agency than a robot or a speaking computer. Thus, a second hypothesis being examined follows: a morally-

charged mistake made by a human agent will still result in a higher perceived trustworthiness toward that agent compared to a machine agent making the same mistake (see Table 1).

“I’m sorry!”—Accounts and Apologies

Although it is useful to understand how humans assign blame to each other as well as machines, and the effect this blame has on perceptions of trustworthiness, it is much more meaningful to explore ways in which moral blame and resulting decrement in perceived trustworthiness might be mitigated. Such information can aid in the design of communication capacities in machine agents. This can also assist in pinpointing areas where machine agents may be more (or less) likely to repair their reputations with humans after making morally repugnant mistakes.

Malle (2016) argues that a key characteristic of morality in robots should be the ability to engage in *moral communication*—including the explanation of any behaviors that violate norms (e.g., morally questionable actions). After all, communication plays a key role in any expression of morality, blame, and forgiveness. To evaluate this, I will explore two methods of communication which can repair damaged reputations and relationships: 1) accounts and 2) apologies. An account can be thought of as a statement made to explain unexpected or unacceptable behavior (Scott & Lyman, 1968). On the other hand, an apology is simply a regretful acknowledgment of the behavior (e.g., “I’m sorry”). Accounts and apologies often occur together, with people offering excuses for their behaviors as part of their apologies. However, here I will examine these phenomena both singly and in conjunction in order to understand their utility when originating from either human or machine agents. Next, I will

discuss how accounts and apologies can be used to repair reputations in humans and how each may aid in the mitigation of blame toward both human and machine agents.

Accounts

Traditionally, accounts have been explored as *vocabularies of motive* (Mills, 1940) and the *grammar of motives* (Burke, 1945). They were labeled this way for their role in explaining human *intent* behind actions; as they are often given to explain one's intent behind unacceptable behavior, including both accidental and intentional wrongdoing. Accounts are often classified as either *excuses* or *justifications* (Scott & Lyman, 1968). Classified in this way, *excuses* are statements made to mitigate responsibility for an act by deflecting blame onto another causal source in the situation. For example, a soldier who accidentally shoots his friend might give the excuse that his friend jumped into the line of fire at the last second. On the other hand, justifications are statements made which take responsibility for an act without accepting moral blame for them. For example, a soldier who shoots an enemy for his country might explain that he did shoot the enemy, as it was his job to do it (i.e., taking responsibility for the behavior, but pointing out that the behavior was required and just).

Accounts are used to appease people in many situations. These contexts range from business interactions to intimate relationships when people may be experiencing doubt about an agent's moral or ethical standing. While both excuses and justifications are used in repairing reputations and relationships, they are not equally effective. For example, in multiple studies, McGraw found that justifications were more effective at repairing the reputations of politicians by constituents, though excuses used in cases of mitigating circumstances could also be effective

(McGraw, 1990, 1991). However, McGraw did not find an effect of either type of account on blame. Conlon and Murray (1996) found similar results in a study examining business relations with customers, showing a significant difference in effect between justifications and excuses. Reasons for their findings center largely on the concept of *taking responsibility*. Ultimately, it is through the admittance of responsibility that human agents can win back satisfaction and support from constituents and customers. But do the same principles apply to machine agents?

Whether or not humans accept accounts given by machine agents may depend on human perception of responsibility in such agents. Just as an agent must have control over itself to be worthy of blame, it should have control to be able to take responsibility. As previously discussed, machine agents don't currently possess a level of control on par with humans. However, when humans perceive machine agents as having more autonomy and independence, perceptions of responsibility will change. For example, a recent study showed that robots displaying lack of effort in a task were judged as having more agency, and potentially more moral responsibility, than identical robots displaying lack of ability (Woerdt & Haselager, 2016). This suggests that machine agents are judged quite differently on the same behavior (not completing a task) based on human perceptions of their abilities. In this way, accounts given by machine agents may be taken very seriously if such agents appear to have the autonomy to take responsibility.

Additional research into this phenomenon has shown that, not only will people attribute more blame to more autonomous robots (i.e., robots capable of acting with little human intervention), but accounts given by such robots can mitigate blame to a greater degree than accounts given by robots perceived as less autonomous (i.e., robots acting with a greater need for human intervention; Kim & Hinds, 2006). The same might be suggested for differences between

machine agents and human agents if one considers human agents to be more autonomous than their machine counterparts. This possibility leads to my third and fourth hypotheses (see Table 1). Specifically, when committing a morally-charged mistake, agents who give an account for the mistake will receive less moral blame than those who do not. Furthermore, when giving the same account, moral blame toward human agents will be mitigated to a greater degree than such blame toward machine agents. My fifth and sixth hypotheses expand this logic with perceived trustworthiness (see Table 1). Specifically, agents who give an account will be perceived as more trustworthy than those who don't. Furthermore, when giving the same account, human agents will be perceived as more trustworthy than machine agents.

Apologies

From an etymological standpoint, apologies are accounts—statements made in defense of oneself (Harper, 2017). However, for the present discussion, the term *apology* will be used independently from the term *account* to refer only to expressions of regret. Thus, it does not include additional statements made in one's defense. The expression of regret offers a useful, albeit simple, alternative to accounts when trying to repair a damaged reputation or relationship. Apologies and accounts share in common the assumption of responsibility, but apologies are further distinguished by expressions of regret, acknowledgment of offense to the victim, and acknowledgment that the victim should be spared the mistake that has occurred (Kort, 1975).

Apologies are used in a wide variety of contexts and can be helpful in mitigating blame and repairing reputations in humans. For example, research has shown that human agents who offer apologies receive less blame than those who do not (Darby & Schlenker, 1982). Likewise,

victims who receive apologies are less likely to punish the offender (Brown & Levinson, 1987). However, the degree to which blame is mitigated and forgiveness is given may depend on the severity of the wrongful behavior; some behaviors may require stronger apologies to elicit the same effects (Slocum, Allan, & Allan, 2011). This suggests that apologies should be carefully crafted to align with the severity of mistake made, especially since over- and under-apologizing may be perceived as insincere.

Apologies may also be useful to mitigate blame and negativity when machine agents make mistakes. For example, in one study examining robot agents that experience a breakdown in service being provided to a customer, apologies were shown to help mitigate the negative effects of their breakdowns—especially for people who desire to maintain a good social relationship with the company (Lee, Kielser, Forlizzi, Srinivasa, & Rybski, 2010). On the other hand, Kaniarasu and Steinfeld (2014) found that excessive apologies by machines can reduce trust toward those machines. Whether this is due to the actual apologies or the admittance of mistakes and self-blame was unclear. Rationally, then, as long as the quantity and quality of the apology is appropriate, it can mitigate blame toward both human and machine agents. What remains to be known is whether such blame is mitigated to a greater degree in human or machine agents. It is possible that, in line with the effect of accounts, apologies are viewed as more effective coming from entities that are “more autonomous” (such as humans). Thus follows my seventh and eighth hypotheses (see Table 1). Specifically, when committing a morally-charged mistake, agents who give an apology for the mistake will receive less moral blame than those who do not. Furthermore, when giving the same apology, moral blame toward human agents will be mitigated to a greater degree than such blame toward machine agents. My ninth and tenth

hypotheses expand this logic with perceived trustworthiness (see Table 1). Specifically, agents who give an apology will be perceived as more trustworthy than those who don't. Furthermore, when giving the same apology, human agents will be perceived as more trustworthy than machine agents.

Either, Or, or Both?

The findings from Kaniarasu and Steinfeld (2014) regarding the issue of over-apology raises the question of whether apologies from machine agents operate as double-edged swords—especially if it leaves humans wondering if the machine is simply incapable of doing its job effectively. This may be especially true if a machine agent is unable to (or simply doesn't) explain its mistake. After all, apologizing for a mistake doesn't let the recipient know that an agent knows *what* was done incorrectly, or give any guarantee that the agent will learn from the mistake and refrain from doing it again.

With this in mind, it is also possible apologies will work more effectively in tandem with accounts. For example, Conlon and Murray (1996) found that complainants who received both accounts and apologies from a company received them quite favorably, making it a useful strategy for public relations. However, other research has shown that combining accounts and apologies isn't always effective. In particular, if an account shows that a wrongful behavior was intentional (i.e., intent to harm or treat wrongly), apologies may do little to mitigate blame (Struthers, Eaton, Santelli, Uchiyama, & Shirvani, 2008). Thus, the combination of accounts and apologies can be expected to work best in cases where a mistake is committed for reasons that don't involve intent to harm. These conclusions lead to my eleventh and twelfth hypotheses (see

Table 1). Agents giving a joint account and apology will receive less moral blame than those who do not. Furthermore, when giving the same joint account and apology, moral blame will be mitigated more for human agents than for machine agents. My final two hypotheses expand prior logic to perceived trustworthiness (see Table 1). Specifically, agents giving a joint account and apology will be perceived as more trustworthy than those who do not. Furthermore, when giving the same joint account and apology, perceived trustworthiness will be greater for human agents than machine agents.

Table 1. List of hypotheses divided by each primary dependent variable of interest.

<u>Moral Blame</u>	<u>Trustworthiness</u>
Hypothesis 1: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in a higher attribution of moral blame compared to a machine agent making the same mistake.	Hypothesis 2: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in higher perceived trustworthiness compared to a machine agent making the same mistake.
Hypothesis 3: There will be a main effect of account, such that agents who give an account for a moral mistake will receive less blame than those who do not.	Hypothesis 5: There will be a main effect of account, such that agents who give an account for a moral mistake will be perceived as more trustworthy than those who do not.
Hypothesis 4: There will be an interaction between agent type and account, such that the effect of account on blame will be qualified by an interaction with agent type.	Hypothesis 6: There will be an interaction between agent type and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with agent type.
Hypothesis 7: There will be a main effect of apology, such that agents who apologize for a moral mistake will receive less blame than those who do not.	Hypothesis 9: There will be a main effect of apology, such that agents who apologize for a moral mistake will be perceived as more trustworthy than those who do not.
Hypothesis 8: There will be an interaction between agent type and apology, such that the effect of apology on blame will be qualified by an interaction with agent type.	Hypothesis 10: There will be an interaction between agent type and apology, such that the effect of apology on perceived trustworthiness will be qualified by an interaction with agent type.
Hypothesis 11: There will be an interaction between apology and account, such that the effect of account on blame will be qualified by an interaction with apology.	Hypothesis 13: There will be an interaction between apology and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology.
Hypothesis 12: There will be an interaction between agent type, apology, and account, such that the effect of account on blame will be qualified by an interaction with apology and agent type.	Hypothesis 14: There will be an interaction between agent type, apology, and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology and agent type.

CHAPTER THREE: EXPERIMENTAL PROCEDURE

“I know I've made some very poor decisions recently, but I can give you my complete assurance that my work will be back to normal.”

- HAL; 2001, A Space Odyssey (Movie)

Design

To expand knowledge on the topics discussed herein, a narrative was designed and implemented as part of a study to examine the role of accounts and apologies in mitigating blame toward human and machine agents. Given the hypotheses presented, a 2x3x3 between subjects factorial design was created, with agent (human vs. machine), account (positive, negative, and neutral valence), and apology (positive, negative, and neutral valence) as the 3 independent variables of interest. The dependent variables were primarily moral blame and trustworthiness, as discussed above. Additionally, information regarding participant sex, displaced blame, and perceived agency were also recorded. Details about the task and materials are discussed next.

Task Overview

Participants were asked to read and respond to the following narrative. The narrative presented a situation in which a citizen hails a ride from a ride-sharing service in order to travel to the airport. The driver of the ride-sharing vehicle runs a stop sign and causes a car accident, which injures the passenger. The narrative begins:

“Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or

the interview. In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.”

The remainder of the narrative was manipulated in order to account for combinations of the three independent variables of interest. Agent type was manipulated as the ride sharing service utilizing either a 1) human, or a 2) machine driver, with participants being informed of the type of agent that was driving. After the agent was introduced, the accident was introduced as such:

“The ride-sharing taxi arrives on time, with {John / an autonomous robot} as its driver.

On the way to the airport, {John / the autonomous robot} runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.”

Account was operationalized as the agent giving an explanation for causing the accident, while also stating its intention (that is, specifically lack of intention to cause harm). Account was manipulated as three categories: 1) the narrative explicitly stating a reason was given (positive valence), 2) the narrative explicitly stating *no* reason was given (negative valence), or 3) the narrative not mentioning reasons at all (neutral valence; control).

- Account given (positive valence): *When John realizes that Charles is injured and will miss his interview, he explains himself, saying “I didn’t intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle.”*

- Account not given (negative valence): *“When John realizes that Charles is injured and will miss his interview, he doesn’t explain himself.”*
- Account control (neutral valence): No additional information beyond the above narrative

Apology was operationalized as the agent expressing regret while giving an admission of wrongdoing for causing the accident. Apology was manipulated as three categories mirroring account: 1) the narrative explicitly stating an apology was given (positive valence), 2) the narrative explicitly stating *no* apology was given (negative valence), or 3) the narrative not mentioning apologies at all (neutral valence; control).

- Apology given (positive valence): *When John realizes that Charles is injured and will miss his interview, he apologizes and explains himself, saying “I regret causing this car accident and your injury. I should have yielded to the other car.”*
- Apology not given (negative valence): *“When John realizes that Charles is injured and will miss his interview, he doesn’t apologize.”*
- Apology control (neutral valence): No additional information beyond the above narrative

The rationale of using control conditions for account and apology is twofold. First, it allowed for experimental control of any priming effect that emerged as a result of the language used in the narrative (i.e., being told “He didn’t apologize” may have a priming effect compared to being told nothing at all). Second, as the goal of the study presented here was to examine apology and account in isolation as well as together, implementing this third category allowed for the exploration of each variable in isolation when combined in a full factorial design.

Due to the factorial nature of this study, the above manipulations resulted in 18 combinations forming the full narratives presented to participants. As such, when recruited, participants were randomly assigned to one of 18 conditions and received that condition exclusively, so as not to cross-contaminate the manipulations made (see *Procedure* for more details). The full narratives used in each condition are included in Appendix A of this document for the reader's perusal.

Materials

The entire study, including the narrative and surveys, was created using Qualtrics, an online service allowing researchers to create complex survey flows and scenarios. The primary dependent variables measured as part of this study included moral blame and trustworthiness. Additional qualitative information was collected concerning participants' displacement of blame (toward others not mentioned in the scenario), as well as their perception of *agency*, or the agent's control of its own behavior. This information was coded and examined to understand general participant attitudes in the study. Finally, demographic information was collected.

Moral Blame

Moral blame was quantified using a 3-item measure of "moral sanctions and praise" developed by Pizarro and colleagues and used in several studies examining morally questionable scenarios (Pizarro et al., 2003; Pizarro Uhlmann, & Solovey, 2003; Pizarro, Laney, Morris, & Loftus, 2006). It should be noted that the modest size of the scale (three items, one per dimension) make it less than ideal for thorough measurement. However, its use in this study

allowed for more direct comparisons of prior studies of moral blame using the same operationalization and quantification of used here. The questions were presented as:

- Please indicate your assessment of Watkins behavior on a scale from 1 to 7 according to the following statements:
 - How moral or immoral was {John's / the autonomous robot driver's} mistake?
 - How blameworthy or praiseworthy is {John / the autonomous robot driver} for the mistake?
 - How positively or negatively should {John / the autonomous robot driver} be judged?

Perceived Trustworthiness

Perceived trustworthiness was quantified using the trustworthiness dimensions (ability, benevolence, and integrity) of Mayer and Davis's (1999) trust scale. In line with Becerra and Gupta's (2003) work, the scale was abbreviated to include the most relevant questions for the study. The questions were slightly modified to fit the narrative participants read, and to elicit participants' attitudes regarding the agent as opposed to their beliefs or behaviors (Fishbein & Ajzen, 1977). The questions were presented as:

- Please indicate the extent to which you agree with the following statements using the scale from one to seven below (Anchors: strongly disagree-neutral-strongly agree):
 - Ability dimension items:
 - I feel {John / the autonomous robot driver} is very capable of performing the job.

- I feel very confident about {John's / the autonomous robot driver's} skills.
- I feel {John / the autonomous robot driver} has much knowledge about the work needing done.
- Benevolence dimension items:
 - I feel {John / the autonomous robot driver} really looks out for what is important for patients.
 - I feel {John / the autonomous robot driver} is very concerned about passengers' welfare.
 - I feel passengers' needs and desires are very important to {John / the autonomous robot driver}.
 - I feel {John / the autonomous robot driver} will go out of the way to help passengers
- Integrity dimension items:
 - I feel {John / the autonomous robot driver} has a strong sense of justice.
 - I feel I never have to worry about whether {John / the autonomous robot driver} will keep a promise.
 - I feel {John / the autonomous robot driver} tries hard to be fair in dealings with others.

Displaced Blame

Displacement of blame was measured first as a yes or no question, followed by an open-ended question to determine if there are others involved in the scenario who participants believed deserved blame for the agent's mistake. The purpose of leaving this question open-ended was to

avoid priming participants that any specific party is blameworthy. In this way, it was possible to gain more automatic reactions from participants on who else (programmers, the company, managers, etc.) may be culpable for the agent's mistake. The question was presented as:

- Is there anyone else who you believe deserves blame for {John's / the autonomous robot driver's} behavior?
 - Yes
 - No
- If yes, who do you believe deserves blame and why?
 - [Participant entered response in a text box]

Perceived Agency

Perceived agency was measured with a single-question on a 5-point Likert scale. The purpose of the question was to track any differences in perception of human and machine agents' autonomy and ability to behave as independent entities. The question was presented as:

- Do you believe {John / the autonomous robot} was in control of {his / its} own behavior?

Participants

Recruitment

Participants were recruited using Amazon Mechanical Turk (MTurk), an online crowdsourcing marketplace. While Mechanical Turk was invented for individuals to complete simple tasks that machines are yet unable to do (e.g., identifying "best" pictures from a group; writing product descriptions), it has become popular as a database for psychological studies. MTurk has the advantage of offering a wider range of participant ages and ethnicities from

around the world (Horton, Rand, & Zeckhauser, 2011). It additionally allows for the control of participant payment to only those who perform appropriately according to pre-set metrics (for example, answering an attention check question correctly), allowing researchers to easily eliminate dubious data.

Research has shown that data collected on MTurk is reliable (Holden, Dennie, & Hicks, 2013) as well as statistically equivalent to data collected at universities and in organizations—as long as language barriers are taken into account (Feitosa, Joseph, & Newman, 2015). That is, studies given in English should only be given to those from populations where English is the primary language. It has thus become common to use MTurk to collect psychological data, particularly in social psychology. This approach has been particularly successful for examining issues of moral judgment (Everett, Pizarro, & Crockett, 2016; Paxton, Ungar, & Greene, 2012) and moral blame (Inbar et al., 2012), which are topics important to this dissertation.

Sample

According to G*Power 3.1, in order to achieve a medium effect size (0.25) and high power (0.95) for a study with 18 conditions, 486 participants were needed. To account for participant drop-out and potentially disqualifying participant responses, extra participants were recruited, totaling 566 adult participants from the USA. Of the 566 participants recruited, 18 participants were removed for failing to complete the study, and an additional 10 participants were excluded for failing attention check questions included in the study. Thirty-one participants were excluded for either leaving the study open too long (>15 minutes) to ensure active participation, or finishing it too quickly (< 90 seconds) to ensure adequate reading

comprehension. Thus, the final sample analyzed was 507. An examination of participant self-reported sex showed that 56.8% of participants were male, 42.6% of participants were female, and 0.6% were “other”. This breakdown is consistent with other demographics data from MTurk (Feitosa et al., 2015).

Procedure

Participants first viewed an IRB-approved informed consent document which included details about the study and what is expected of participants (see Appendix B). Once participants agreed to participate in the study, they were randomly assigned to one of 18 conditions, and presented with the narrative discussed in the *Task Overview* above (see Appendix A for specific narrative breakdowns). After reading the full narrative, participants completed post-task questionnaires. A “test question” was included to check for participants not reading the questions and simply answering to get through. Once the post-task questionnaires were complete, participants were thanked and paid \$0.15 for their participation per MTurk’s standards of pay.

CHAPTER FOUR: EXPERIMENTAL RESULTS

“Never ruin an apology with an excuse.”

- Benjamin Franklin

IBM SPSS Statistics, version 23.0, was used to complete all analyses, with $\alpha = .05$ used at the criterion for significance, and η_p^2 examined for effects. Analyses began with an examination of survey conditions. Due to elimination of participants, sample sizes between the 18 conditions were not identical, but were fairly close, ranging from 26 to 30 participants per condition. A further examination was made of sex (male, female, and other) breakdown per condition. While the ratio of male (56.8%) to female (42.6%) in the overall sample matched that of MTurk sampling demographics (Feitosa et al., 2015), closer examination of the data showed that the ratio of male to female participants was not upheld evenly between conditions. In particular, conditions 8 (69.2% male) and 14 (76.7% male) deviated greatly from the standard ratio. In order to account for any systematic differences that might arise due to sex between conditions, sex was included an independent variable in all analyses. An examination of power using G*Power 3.1 showed that this could be done with the current sample size ($N = 507$) while achieving adequate power (.85).

Moral Blame

An examination of the moral sanctions and praise index developed by Pizarro and colleagues (2003) showed very modest inter-item correlations, and Cronbach’s $\alpha = .69$. While this value leaves some doubt to the reliability of the scale, the items were nevertheless combined and analyzed as a single scale in order to make comparisons to prior work on moral blame. A

3x2x3x3 Analysis of Variance (ANOVA) was completed with the following independent variables: sex (male, female, other), agent (human, machine), apology (positive valence, negative valence, control), and account (positive valence, negative valence, control) and the combined blame score as the dependent variable. All assumptions for ANOVA were upheld, so results were analyzed as originally intended, with a focus on the primary hypotheses of interest. However, any significant findings regarding sex differences were noted as well.

Sex Differences

A significant main effect of sex was observed, $F(2, 468) = 3.06, p < .05, \eta_p^2 = 0.01$, with females generally scoring significantly ($p < .05$) higher in moral blame ($\mu = 5.61, SE = .06$) than males ($\mu = 5.42, SE = .05$). Furthermore, there was an interaction effect between sex and agent type, $F(2, 468) = 3.28, p < .05, \eta_p^2 = .01$, suggesting that males and females judged the human and machine agents differently (see Figure 3). When looking at the graph, it should also be noted that the extremely small group ($N = 3$) of “other” sex respondents shows what appears to be an additional interaction. However, the “other” group is too small for any conclusions to be made.

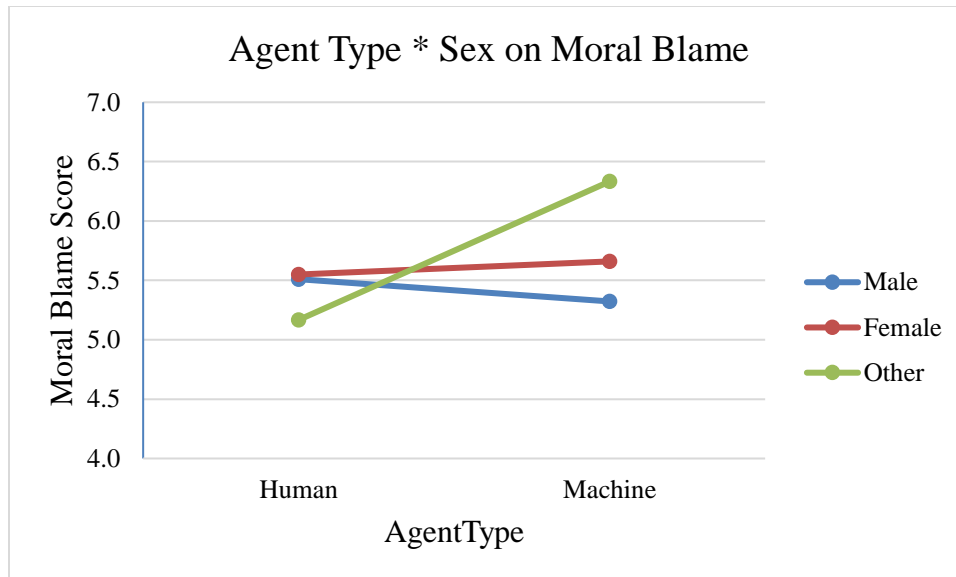


Figure 3. Estimated marginal means for moral blame for sex across agent type show an interaction wherein males, females, and other sex assign different levels of blame to different types of agents. Post hoc comparisons showed a significant difference between males' and females' assignments of blame.

Examination of Hypotheses

An examination of the main effects and interactions hypothesized prior to the study showed that hypotheses 3, 7, and 11 regarding moral blame were upheld (see Table 2). While all findings and effect sizes are noted in the table, significant findings were examined in more detail with post hoc pairwise comparisons, as discussed next.

Hypothesis 3: A main effect of account was found, $F(2, 468) = 4.35, p < .05, \eta_p^2 = .02$, with the control conditions (neutral valence) receiving higher blame scores ($\mu = 5.62, SE = .07$) than negative valence ($\mu = 5.53, SE = .11$) and positive valence ($\mu = 5.4, SE = .09$) conditions (see Figure 4). This suggests that participants assigned greater blame when no mention of an account was made at all. However, post hoc pairwise comparisons showed no significant differences between these respective conditions.

Hypothesis 7: A main effect of apology was found, $F(2, 468) = 4.21, p < .05, \eta_p^2 = .02$.

Post hoc comparisons showed that there was a significant difference in blame scores between the negative valence and the neutral valence conditions ($p < .05$), as well as the negative valence and positive valence conditions ($p < .05$), with blame scores in the negative valence conditions being higher ($\mu = 5.70, SE = .09$) than the neutral ($\mu = 5.43, SE = .07$) and positive ($\mu = 5.42, SE = .11$) conditions, respectively (see Figure 4). This suggests that participants were more likely to assign blame when it was noted that the agent did *not* apologize. Indeed, there seems to be very little difference in blame scores when the agent apologized compared to when no mention of apology was made. However, when it was directly stated that the agent did *not* apologize, participants assigned blame.

Hypothesis 11: An interaction between apology and account was found $F(4, 468) = 2.60, p < .05, \eta_p^2 = .02$, suggesting that the presence of an apology enhanced the utility of an account for mitigating blame toward agents (see Figure 4). Specifically, combining neutral apology valence with positive account valence resulted in the lowest level of blame whereas combining negative account and apology valences resulted in the highest blame.

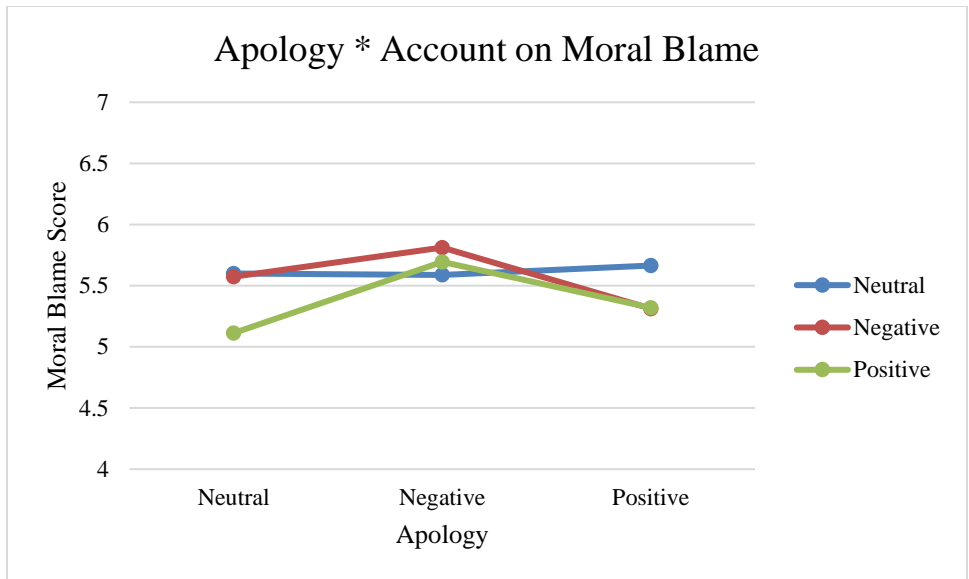


Figure 4. Estimated marginal means for moral blame for account across apology show an interaction between account and apology, wherein coupling a neutral apology valence with a positive account valence resulted in the lowest level of moral blame.

Table 2. Hypotheses regarding moral blame and their respective outcomes.

<u>Moral Blame Hypotheses</u>	<u>Outcomes</u>
Hypothesis 1: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in a higher attribution of moral blame compared to a machine agent making the same mistake.	Not upheld; $F(1, 468) = 2.06, p = .15, \eta_p^2 = .004$
Hypothesis 3: There will be a main effect of account, such that agents who give an account for a moral mistake will receive less blame than those who do not.	Upheld; $F(2, 468) = 4.35, p < .05, \eta_p^2 = .02$
Hypothesis 4: There will be an interaction between agent type and account, such that the effect of account on blame will be qualified by an interaction with agent type.	Not upheld; $F(2, 468) = 1.23, p = .29, \eta_p^2 = .01$
Hypothesis 7: There will be a main effect of apology, such that agents who apologize for a moral mistake will receive less blame than those who do not.	Partially upheld; $F(2, 468) = 4.21, p < .05, \eta_p^2 = .02$
Hypothesis 8: There will be an interaction between agent type and apology, such that the effect of apology on blame will be qualified by an interaction with agent type.	Not upheld; $F(2, 468) = 0.68, p = .51, \eta_p^2 = .003$
Hypothesis 11: There will be an interaction between apology and account, such that the effect of account on blame will be qualified by an interaction with apology.	Partially upheld; $F(4, 468) = 2.60, p < .05, \eta_p^2 = .02$
Hypothesis 12: There will be an interaction between agent type, apology, and account, such that the effect of account on blame will be qualified by an interaction with apology and agent type.	Not upheld; $F(4, 468) = 1.01, p = .40, \eta_p^2 = .01$

Perceived Trustworthiness

An examination of three subscales of the trustworthiness index developed by Mayer and colleagues (1995) showed very strong inter-item correlations and Cronbach's $\alpha = .86$. Thus, in keeping with the original intent of the scale, items were initially combined and analyzed as a single scale in order to make comparisons to prior work on moral blame. After this primary

analysis, an additional analysis was completed to examine the trustworthiness dimensions individually. To start, a 3x2x3x3 ANOVA was completed with the following independent variables: sex (male, female, other), agent (human, machine), apology (positive valence, negative valence, control), and account (positive valence, negative valence, control) and the combined trustworthiness score as the dependent variable. All assumptions for ANOVA were upheld, so results were analyzed as originally intended, with a focus on the primary hypotheses of interest. However, any significant findings regarding sex differences were noted as well.

Sex Differences

A significant main effect of sex was found, $F(2, 468) = 6.14, p < .05, \eta_p^2 = .03$, with males generally ranking the agents as more trustworthy ($\mu = 2.97, SE = .07$) than females did ($\mu = 2.62, SE = .08$).

Examination of Hypotheses

An examination of the main effects and interactions hypothesized prior to the study showed that hypotheses 5, 9, and 10 regarding moral blame were upheld (see Table 3). While all findings and effect sizes are noted in the table, significant findings were examined in more detail with post hoc pairwise comparisons, as discussed next.

Hypothesis 5: A main effect of account was found, $F(2, 468) = 9.80, p < .001, \eta_p^2 = .04$. Post hoc comparisons showed a significant difference between control conditions (neutral valence) and positive valence conditions ($p < .05$), and between positive valence conditions and negative valence conditions ($p < .05$). Specifically, trustworthiness scores were higher in the positive valence conditions ($\mu = 3.03, SE = .12$), than the negative valence ($\mu = 2.65, SE = .14$)

and neutral valence conditions ($\mu = 2.63$, $SE = .09$). This suggests participants found agents more trustworthy when the agents gave an account for their behavior (see Figure 5).

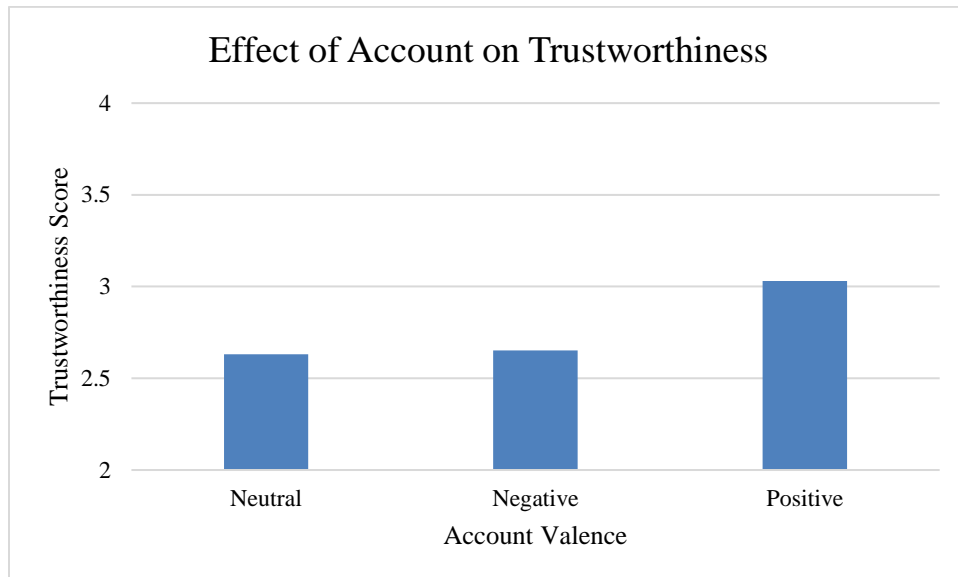


Figure 5. Estimated marginal means reflecting the main effect of account on perceived trustworthiness.

Hypothesis 9: A main effect of apology was found, $F(2, 468) = 23.10$, $p < .001$, $\eta_p^2 = .09$. Post hoc comparisons showed a significant difference between control conditions (neutral valence) and negative valence conditions ($p < .05$), and between positive valence conditions and negative valence conditions ($p < .001$). Specifically, trustworthiness scores were higher in the positive valence conditions ($\mu = 3.13$, $SE = .14$), than the neutral valence ($\mu = 2.83$, $SE = .09$), and negative valence conditions ($\mu = 2.34$, $SE = .12$). This suggests participants found agents more trustworthy when the agents gave an apology for their behavior (see Figure 6).

Hypothesis 10: An interaction effect was found for agent type and apology, $F(2, 468) = 3.41$, $p < .05$, $\eta_p^2 = .01$, suggesting that the effect of the apology on trustworthiness was determined in part by the type of agent giving the apology. As can be seen in Figure 6, rated

trustworthiness was affected much more by apologies given in the human agent condition than in the machine agent condition. This may be due to a difference in perceptions of agency between human and machine agents (discussed below).

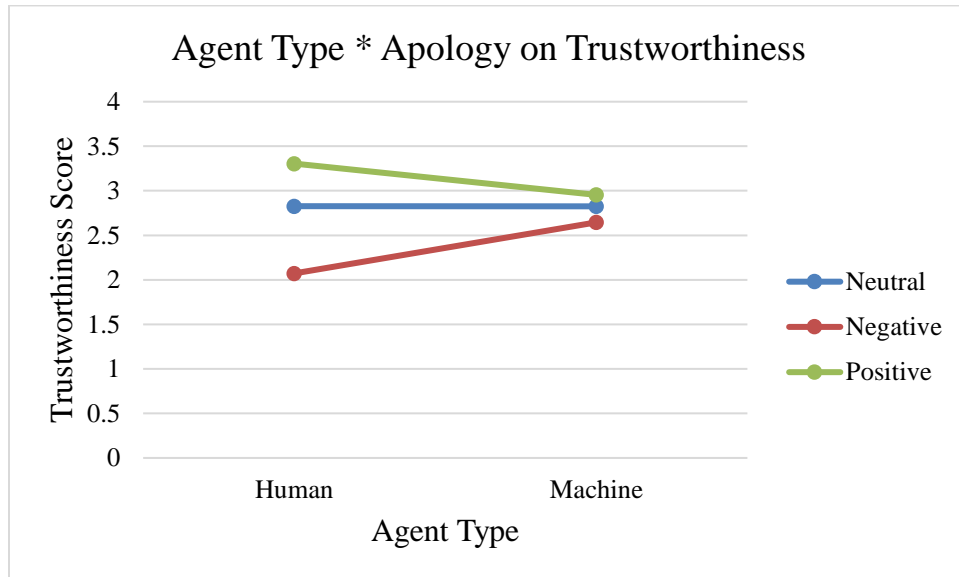


Figure 6. Estimated marginal means reflect an interaction effect between agent type and apology on trustworthiness scores. Visual examination of the spread of scores for each agent type shows a much more consistent ranking of trustworthiness of the machine agent, whereas negative and positive valence of apology resulted in quite disparate rankings of trustworthiness toward the human agent.

Table 3. Hypotheses regarding trustworthiness and their respective outcomes.

<u>Trustworthiness Hypotheses</u>	<u>Outcomes</u>
Hypothesis 2: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in higher perceived trustworthiness compared to a machine agent making the same mistake.	Not upheld; $F(1, 468) = .83, p = .36, \eta_p^2 = .002$
Hypothesis 5: There will be a main effect of account, such that agents who give an account for a moral mistake will be perceived as more trustworthy than those who do not.	Upheld; $F(2, 468) = 9.80, p < .001, \eta_p^2 = .04$
Hypothesis 6: There will be an interaction between agent type and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with agent type.	Not upheld; $F(2, 468) = 1.28, p = .28, \eta_p^2 = .01$
Hypothesis 9: There will be a main effect of apology, such that agents who apologize for a moral mistake will be perceived as more trustworthy than those who do not.	Upheld; $F(2, 468) = 23.10, p < .001, \eta_p^2 = .09$
Hypothesis 10: There will be an interaction between agent type and apology, such that the effect of apology on perceived trustworthiness will be qualified by an interaction with agent type.	Upheld; $F(2, 468) = 3.41, p < .05, \eta_p^2 = .01$
Hypothesis 13: There will be an interaction between apology and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology.	Not upheld; $F(4, 468) = .49, p = .74, \eta_p^2 = .004$
Hypothesis 14: There will be an interaction between agent type, apology, and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology and agent type.	Not upheld; $F(4, 468) = .99, p = .41, \eta_p^2 = .01$

Trustworthiness Dimensions

In order to gain a better understanding of participants' perception of ability, benevolence, and integrity of the agents, a 3x2x3x3 Multivariate Analysis of Variance (MANOVA) was completed using the 3 subscales that formed the trustworthiness survey as the dependent

variables of interest. The independent variables were identical to the prior ANOVAs. All assumptions for ANOVA were upheld, except Box's M, which is known to be particularly sensitive to large sample sizes and complex analyses. As such, the results were interpreted using Pillai's Trace.

Sex Differences

Using Pillai's trace, a multivariate effect of sex was found, $V = .04$, $F(6, 934) = 2.77$, $p < .05$, $\eta_p^2 = 0.02$. Specifically, there was a main effect of sex on the *ability* and *benevolence* dimensions of the trustworthiness scale; $F(2, 468) = 5.31$, $p < .05$, $\eta_p^2 = 0.02$, and $F(2, 468) = p < .05$, $\eta_p^2 = .03$, respectively. In general, men rated the agents as being higher in *ability* ($\mu = 3.02$, $SE = .08$) and *benevolence* ($\mu = 2.79$, $SE = .08$) compared to women ($\mu = 2.65$, $SE = .10$ and $\mu = 2.35$, $SE = .09$, respectively). Additionally, an interaction effect between sex and account emerged, $F(2, 468) = 3.19$, $p < .05$, $\eta_p^2 = .01$, suggesting that men and women judged the humans and agents differently.

Examination of Primary Variables

Using Pillai's trace, a multivariate effect of account was found, $V = .05$, $F(6, 934) = 4.1$, $p < .001$, $\eta_p^2 = .03$. Specifically, there was a main effect of account on all 3 subscales: *ability*, $F(2, 468) = 3.79$, $p < .05$, $\eta_p^2 = .02$; *benevolence*, $F(2, 468) = 10.84$, $p < .001$, $\eta_p^2 = .04$; and *integrity*, $F(2, 468) = 9.41$, $p < .001$, $\eta_p^2 = .04$. Across all three subscales, participants rated agents more highly when an account was given (see Table 4).

Table 4. Means and SE of participant ratings of ability, benevolence, and integrity of agents, broken down by *account* given. Note that the means presented here are based on modified population marginal mean.

Dependent Variable	Valence	Mean	Std. Error
Trustworthiness: Ability	Neutral	2.681	.107
	Negative	2.743	.169
	Positive	2.939	.144
Trustworthiness: Benevolence	Neutral	2.401	.102
	Negative	2.384	.161
	Positive	2.901	.137
Trustworthiness: Integrity	Neutral	2.812	.091
	Negative	2.823	.143
	Positive	3.253	.122

Using Pillai's trace, a multivariate effect of apology was found, $V = .12$, $F(6, 934) = 9.84$, $p < .001$, $\eta_p^2 = .06$. Specifically, there was a main effect of apology on all 3 subscales of trustworthiness: *ability*, $F(2, 436) = 7.57$, $p < .05$, $\eta_p^2 = .03$; *benevolence*, $F(2, 436) = 20.25$, $p < .001$, $\eta_p^2 = .08$; and *integrity*, $F(2, 436) = 30.33$, $p < .001$, $\eta_p^2 = .12$. Across all three of these subscales, participants rated agents more highly when an apology was given (see Table 5).

Table 5. Means and SE of participant ratings of ability, benevolence, and integrity of agents, broken down by *apology* given. Note that the means presented here are based on modified population marginal mean.

Dependent Variable	Valence	Mean	Std. Error
Trustworthiness: Ability	Neutral	2.856	.110
	Negative	2.438	.146
	Positive	3.058	.166
Trustworthiness: Benevolence	Neutral	2.602	.105
	Negative	2.100	.139
	Positive	2.955	.158
Trustworthiness: Integrity	Neutral	3.014	.093
	Negative	2.472	.124
	Positive	3.375	.141

Displaced Blame and Agency

Displaced Blame

In addition to the primary analyses on moral blame, an examination of displaced blame was completed in order to gain a more thorough understanding of the nature of blame toward human and machine agents. As suggested by the lack of a primary effect of agent type, the sample as a whole did not vary widely in blame toward human and machine agents. However, when asked if there were any *other* parties to blame in the car accident presented to them, participants responded very differently in the human and machine agent conditions. While only a few participants agreed someone else was to blame in the human conditions ($N = 17$, or 6.8%), many responded quite the opposite in machine agent conditions ($N = 164$, or 63.5%).

An additional question asking participants to identify *who* else they believed to be deserving of blame identified several parties to consider, especially in the machine agent conditions. Coding of the free responses resulted in six categories of *blame targets* identified by participants (see Table 6). While no single target far surpassed others in the human agent conditions, participants were most likely to choose the agent's creator (i.e., programmer, designer, or company that created the machine) as the secondary source of blame in the machine agent conditions. Note that no further information was gathered from participants, so it is unknown just how much blame participants put toward the machine's creators compared to the machine driver.

Table 6. Coded categories of targets for displaced blame. Participants answered the “displaced blame” question organically, and the categories were created based on trends which emerged from the responses.

Displaced Blame Category	Frequency	Percentage
Creator	137	75.7%
The other driver	13	7.18%
Charles	10	5.52%
Other	9	4.97%
Company	6	3.32%
Several Parties	6	3.31%

Perceived Agency

An examination of perceived agency was completed to understand if participants viewed the human agent as being more in control of its behavior than the machine agent. As noted by the lack of a primary effect of agent type, participants did not vary widely in their perceived trustworthiness of the human and machine agents, nor did they vary widely in their perceptions of the agents’ ability, benevolence, or integrity. However, when asked if they believed the agent was in control of its behavior, participants appeared much more likely to answer “definitely yes” with the human agent, while their responses were quite varied with the machine agent (see Figure 7).

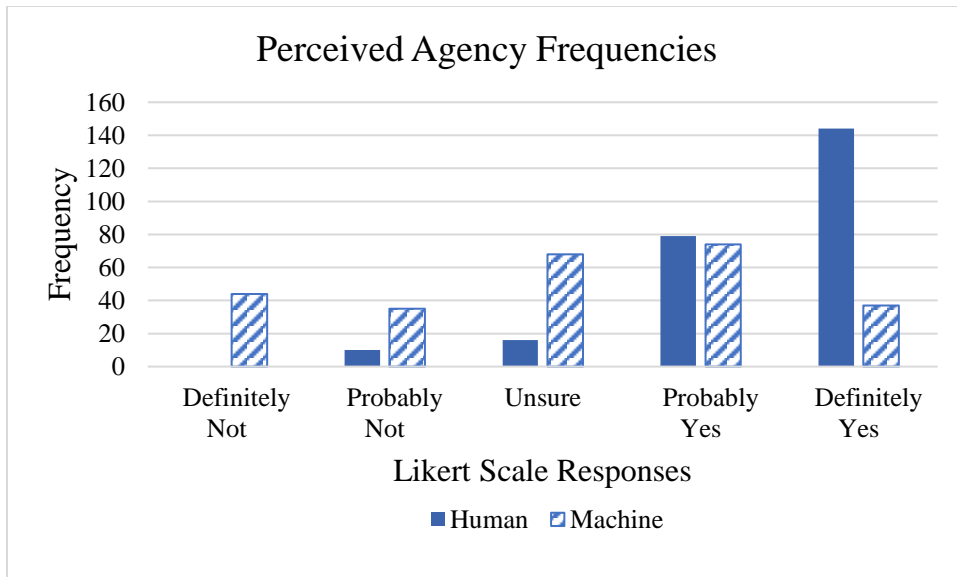


Figure 7. Frequencies reflecting participants’ perception of agency. Specifically, participants were asked if they believed the (human or machine) agent was in control of its behavior. As is reflected in this graph, people were much more likely to respond “definitely yes” for the human condition, but “probably yes” or “unsure” for the machine condition.

CHAPTER FIVE: DISCUSSION

“Trust me.”

- The Terminator; Terminator 3: Rise of the Machines (Movie)

Table 7 shows that support was found for several hypotheses, though perhaps not for those most compelling ones regarding differences in attitudes between human and machine agents. Still, there is a great deal to learn from the hypotheses supported, those not supported, as well as the additional analyses completed.

Table 7. Review of hypotheses, including whether the study presented here lends support for these hypotheses (grey boxes), or not (white boxes).

<u>Moral Blame</u>	<u>Trustworthiness</u>
Hypothesis 1: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in a higher attribution of moral blame compared to a machine agent making the same mistake.	Hypothesis 2: There will be a main effect of agent type, such that a moral mistake made by a human agent will result in higher perceived trustworthiness compared to a machine agent making the same mistake.
Hypothesis 3, Upheld: There will be a main effect of account, such that agents who give an account for a moral mistake will receive less blame than those who do not.	Hypothesis 5, Upheld: There will be a main effect of account, such that agents who give an account for a moral mistake will be perceived as more trustworthy than those who do not.
Hypothesis 4: There will be an interaction between agent type and account, such that the effect of account on blame will be qualified by an interaction with agent type.	Hypothesis 6: There will be an interaction between agent type and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with agent type.
Hypothesis 7, Partially Upheld: There will be a main effect of apology, such that agents who apologize for a moral mistake will receive less blame than those who do not.	Hypothesis 9, Upheld: There will be a main effect of apology, such that agents who apologize for a moral mistake will be perceived as more trustworthy than those who do not.
Hypothesis 8: There will be an interaction between agent type and apology, such that the effect of apology on blame will be qualified by an interaction with agent type.	Hypothesis 10, Upheld: There will be an interaction between agent type and apology, such that the effect of apology on perceived trustworthiness will be qualified by an interaction with agent type.
Hypothesis 11, Partially Upheld: There will be an interaction between apology and account, such that the effect of account on blame will be qualified by an interaction with apology.	Hypothesis 13: There will be an interaction between apology and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology.
Hypothesis 12: There will be an interaction between agent type, apology, and account, such that the effect of account on blame will be qualified by an interaction with apology and agent type.	Hypothesis 14: There will be an interaction between agent type, apology, and account, such that the effect of account on perceived trustworthiness will be qualified by an interaction with apology and agent type.

How Do I Blame Thee?

Hypothesis 1 proposed that participants' assignment of moral blame would be affected by the type of agent committing the mistake in the narrative. Specifically, it was hypothesized that

human agents would receive more blame due to expectations of greater autonomy in humans compared to intelligent machines. The results of this study indicated that such an effect is lacking. However, an interaction effect between sex and agent type showed that males and females judged the human and machine agents differently, suggesting further examination may be needed in order to understand under which demographic circumstances agent type matters (more on this in *Yes, Sex Matters*).

An interesting contrast to the lack of main effect of agent type is the high rate of displaced blame found in participant responses to the study. As noted in the results, 63.5% of participants in the machine conditions stated that they believed another party was additionally to blame for the car accident caused by the machine agent. Additionally, participants' perception of the machine agent's self-control varied quite a bit in comparison to the human agent, suggesting that many were aware that even a highly intelligent machine agent doesn't have the *mens rea* of a human. With such a high rate of displaced blame and such doubt of the machine's agency, one might expect that moral blame toward the machine agent should be less than it is toward the human agent. Instead, it seems that people are willing to at least initially assign as much blame to machine agents as human agents, with the understanding that additional blame is deserved by the machine agent's creators. Such a process can be likened to the blame directed toward parents when kids make mistakes or pet owners when pets make mistakes. While one feels anger or frustration toward the child or pet, they find the parent or owner ultimately responsible (e.g., Drexler, 2012). This finding isn't at all surprising when you consider public reactions to recent catastrophic events involving intelligent machines; there is always another party believed to be responsible by the public (Ohnsman, 2017).

Hypothesis 3 concerned the utilization of *account* as a moral blame mitigation strategy, with the expectation that moral blame would be lower in conditions which included accounts. While this study provided support for the hypothesis, the specific effect is somewhat unclear. There were no significant differences between categories of account (positive, negative, and neutral valence). Furthermore, an examination of the means showed that being told of an account (positive valence) and being told there was no account (negative valence) elicited very similar results. It was actually in the absence of information altogether (neutral valence) that blame scores were highest. This might suggest that participants were less concerned with the presence of an account than expected. However, no definitive conclusions can be drawn with such an unclear effect.

Hypothesis 7 concerned the utilization of *apology* as a moral blame strategy, with the expectation that moral blame would be lower in conditions which included apologies. Once again, this hypothesis was supported by the findings of the study. However, pairwise comparisons indicated that the reason for this effect was primarily due to increased blame in conditions where it was explicitly stated that *no* apology was given (negative valence). This reflects the concerns raised earlier that the negative valence condition could elicit a priming effect. Indeed, based on this finding, it can't be said that giving an apology is a useful mitigation strategy for moral blame, since it elicits very similar moral blame to being told nothing at all. Instead, the main lesson to be learned here is that when people are alerted to the lack of remorse on the part of an agent, they will assign higher blame than otherwise.

Hypothesis 11 predicted an interaction effect between apology and account, whereby their combination would lead to a further alleviation of blame than when presented alone. This finding was upheld, but not necessarily in the direction predicted. While the presentation of an account and apology certainly led to lower moral blame scores, the lowest blame scores were found when an account was given (positive valence) with no mention of apology (neutral valence). Furthermore, there appeared to be a priming effect of negative valence conditions, where a stated lack of both account and apology led to higher blame scores than any other combination. This once again suggests that a stated lack of remorse and lack of acknowledgement on the part of the agent may lead to higher blame than merely keeping silent.

Hypotheses 4 and 8 concerned interaction effects between agent type and account, and agent type and apology, respectively. Likewise, hypothesis 12 predicted a three-way interaction between agent type, account, and apology. None of these hypotheses were upheld, in line with a lack of main effect of agent type. This suggests that people are likely to respond equally to apologies and accounts given by human and machine agents. This might be promising news for designers of machine agents, as it indicates that certain human communications (e.g., apologies) coming from machine agents may actually be equally as effective as it is coming from humans. However, further examination of this hypothesis is needed, particularly in real-life scenarios with humans interacting with such agents (more on this in *Future Research*).

Are You Worthy of My Trust?

Hypothesis 2 proposed that participants' perceptions of trustworthiness would be affected by the type of agent committing the mistake in the narrative. Specifically, it was expected that

human agents would be perceived as more trustworthy due to expectations of greater autonomy in humans compared to intelligent machines. This was particularly expected given the scale used to measure trustworthiness, as the scale involved dimensions which generally aren't thought to be possessed by machines (i.e., *benevolence* and *integrity*, Mayer et al., 1995). This hypothesis was not upheld. However, hypothesis 10, which suggested an interaction effect between agent type and apology, was upheld, as an explicitly stated lack of apology caused perceived trustworthiness of human agents to be much lower than that of machine agents (more on this below).

Hypothesis 5 suggested that giving an *account* may be a useful strategy for upholding trustworthiness when a moral mistake is made, and the findings in this study support that notion. Specifically, the presence of an account (positive valence) appeared to elicit higher trustworthiness than a stated lack of an account (negative valence) or no mention of it at all (neutral valence). Furthermore, hypothesis 6 was *not* upheld, suggesting that participants were equally likely to be affected by the account given regardless of the agent giving it. This suggests that explaining problematic behaviors may indeed affect the development of perceived trustworthiness between people, as well as between people and machines. This has far-reaching implications for designers who endeavor to improve trust in human-machine interaction. However, what is unknown is just how far this development reaches. This study only examined perceived *trustworthiness*, which is quantifiable as part of one's first impression. However, whether this effect extends to the development of trust itself is unclear and should be examined in this specific context (more on this in *Future Research*). More applied research on interaction between humans and machine agents suggests that trust can be improved by machine agents

explaining their behavior (Chen & Barnes, 2014; Mercado et al., 2016; Stowers et al., 2017). The hope is that this is the case for explaining mistakes as well.

Hypothesis 9 predicted that giving an *apology* may be a useful strategy for upholding trustworthiness when a moral mistake is made, and the findings in this study support that notion. As with accounts, giving an apology (positive valence) appeared to elicit higher trustworthiness than a stated lack of an apology (negative valence) or no mention of it at all (neutral valence). Additionally, hypothesis 10 was upheld, specifically as participants rated the human agent much lower on trustworthiness when an explicit statement of lack of apology was given (negative valence) compared to the machine agent. This suggests that there may be a priming effect of the negative valence conditions and that people are much more skeptical when told human agents have failed to apologize. Reasons for this could be that people have higher expectations of humans—that they *should* apologize, whereas a machine agent apologizing or failing to apologize may be inconsequential in comparison. This certainly aligns with the hypothetical expectation that humans hold each other to a different standard for trustworthiness than they hold machines.

Neither hypothesis 6, suggesting an interaction between account and agent type, hypothesis 13, suggesting an interaction between apology and account, nor hypothesis 14, suggesting an interaction between apology, account, and agent, were upheld. This suggests that combining apology and account was not effective for eliciting higher trustworthiness in either agent. This is an interesting contrast to the effectiveness of combined account and apology in mitigating blame toward human and machine agents (as discussed above). But, more

importantly, this lends support to the idea that giving an account may be equally effective for both human and machine agents (hypothesis 6), something that designers can take heart in when creating explainable AI or *explainable agency*, that is the ability to explain one's own behavior (see Langley, Meadows, Sridharan, & Choi, 2017).

Further examination of the trustworthiness dimensions present interesting details to consider when discussing perception of human and machine agents. While there was no main effect of agent type, there were multivariate and main effects of account and apology. Particularly noteworthy is the fact that, while ratings on all three dimensions increased when account and apology were presented, the dimension rated most highly was integrity. This suggests that, perhaps, giving accounts and apologies on behalf of morally problematic behavior may have particular implications for the establishment, repair, and maintenance of one's integrity. This is consistent with prior research on the effectiveness of accounts and apologies (Wildman, 2011). However, such a claim in the particular context of morality should be examined in more detail (see *Future Research*).

Yes, Sex Matters

Analyzing sex in this study exposed unexpected findings that haven't been thoroughly examined in research on moral blame. Specifically, the findings that females gave higher ratings of moral blame and lower ratings of trustworthiness suggest that females may be quicker to judge and overall more skeptical of both human and machine agents when presented with a mistake that calls into question the morality of the agent in question. This might be particularly important to consider when taking into account that females were even harsher of machine

agents than human agents. Further research should be done to examine this effect in more detail. Furthermore, researchers should consider accounting for sex in future work on moral blame and trustworthiness (more on this in *Future Research*).

Theoretical and Practical Implications

The findings presented herein have several implications in modern society, where the presence of machine agents is increasing rapidly. As machines take on more diverse roles, we as a society will see more instances of machines being responsible for human life as well as for human injury and death. The goal of this study was to examine potential differences in human attitudes toward human and machine agents making the same morally-laden, harmful mistake. Findings presented can be used to inform theories on moral blame and trustworthiness, as well as the practical implementation of machine agents across areas of society.

Theoretical Implications

To date, theories on moral blame focus almost exclusively on human agents. This study has not only contributed to a shift in focus toward machine agents, but has allowed for the direct comparison of attitudes toward both human and machine agents. The findings here can be used to contribute to new theories on moral blame as well as the exploration of additional variables that should be considered in these theories.

The present study's findings on the utility of apology and accounts for improving perception of trustworthiness also has implications for developing theories on trustworthiness and trust in human-machine interaction. The majority of research regarding human-machine interaction has failed to take into account fundamental truths learned from social psychology and

related areas, often leaving a gap in our understanding of the full arsenal of tools available to building relationships between humans and machines. Rather than focusing exclusively on relationships between humans and machines, researchers would be well-advised to begin considering what can be learned from human relationships to improve human-machine relationships. Some have already embarked on this endeavor (Kessler, Stowers, Brill, & Hancock, 2017), but much more can still be done.

Practical Implications

While, generally speaking, no difference was found between agent types, a great deal of information was found regarding mitigation strategies—specifically giving accounts and apologies. In some cases, these strategies weren't found to be effective as much as the stated lack of these behaviors was found to be harmful. However, the general lack of difference in blame toward human and machine agents suggests that accounts and apologies coming from machine agents may be as effective as those coming from human agents. As such, companies should consider whether the implementation of such moral blame mitigation strategies in machine agents with the potential to harm might be an effective public relations strategy for making machines more acceptable to society.

The effectiveness of accounts and apologies is clearer when examining perceived trustworthiness, as conditions stating an apology or an account led to increased levels of trustworthiness. Furthermore, while accounts appeared to be equally beneficial to the perceived trustworthiness of human and machine agents, apologies were particularly helpful for machine agents. Given that trust is a key consideration for budding human-machine relationships

(Hancock et al., 2011; Stowers et al., 2017), this finding may have lasting implications for the design and implementation of machine agents in society. However, this should still be considered alongside prior findings that being *over-apologetic* can be counter-productive to the reputation of machines (Kaniarasu & Steinfeld, 2014). Whereas Kaniarasu and Steinfeld found that being over-apologetic made a machine appear incompetent, what remains unknown is whether the context of the apology partially determines this effect. For example, being over-apologetic for a morally questionable behavior may not be as problematic as doing the same thing for a simple, harmless mistake.

Additional implications can be found in the realm of law and policy. As stated previously, when machine agents (i.e., the self-driving Tesla in Florida; Muoio, 2016) have been involved in the harm of humans, formal inquiries into the origin of blame (e.g., NHTSA, 2017) have been integral to the understanding of the machine's role in the situation. Currently, it appears that public and politico-legal opinions on the matter may not always match. Results from the study presented herein have suggested that the public is not only willing to blame a machine for its mistakes, but may also be interested in seeing that the makers of these machines are held accountable.

Future Research

Several questions remain which were not explored in this study. Specifically, what additional characteristics of people should be examined in relation to moral blame and trustworthiness toward machine agents, especially self-driving vehicles? At present, age, religion, and ethnicity could play a role. However, as machines continue to permeate different

areas of society, we might expect these characteristics to become less defining. Other characteristics, such as gaming experience and technology acceptance, could act as mediators and should additionally be examined for greater understanding of various attitudes toward machine agents.

Future research should also move this examination into a more applied arena. While the findings presented herein are meaningful, they are restricted to the attitudes that people have when faced with purely hypothetical situations. Thus, it is important to determine whether these findings are upheld in laboratory or even real-world interactions with humans and machines making morally-laden mistakes. However, given the ethically questionable nature of creating scenarios centered on harm, researchers must be creative in their research designs and use simpler moral mistakes for laboratory examinations.

Further examination is also required for several hypotheses which were supported in this study. Specifically, it is important to test whether human-like communications (such as apologies) from machine agents are really as effective as they are from human agents. This study showed that apologizing and giving an account had similar effects on moral blame toward human and machine agents. Thus, a re-examination of this hypothesis in another context, particularly a more applied study, or a study with a stronger manipulation, may lead to other findings.

Supplemental analyses completed for the trustworthiness dimensions suggested that integrity may be an area of particular interest when examining human reactions to moral mistakes made by human and machine agents. This is a ripe area of research, especially as factors such as benevolence and integrity of machine agents have not been considered nearly as

much as factors such as ability. Is it possible to increase perceptions of integrity in machine agents? Is it necessary to? Future research should examine these questions.

Finally, more research should be done to examine sex differences in moral blame and trustworthiness research. At the very least, if sex is not included as a variable in studies completed on these topics, it should be controlled for through the use of matched-pairs research designs. Furthermore, practitioners should consider whether research should be done on the design and creation of sex-or-gender-specific machine agents in driving and other contexts.

APPENDIX A: NARRATIVES BY CONDITION

Note: Each condition is labeled as having a positive (y), negative (n), or neutral (o) valence for account and apology. Positive valence indicates the narrative states the account or apology was given. Negative valence indicates that the narrative states no account of apology was given. Neutral valence indicates that the narrative withheld any additional statement.

Condition 1: Account (y) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he apologizes and explains himself, saying “I regret causing this car accident and your injury. I should have yielded to the other car. I didn’t intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle.”

Condition 2: Account (n) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he doesn't explain himself. He apologizes, saying "I regret causing this car accident and your injury. I should have yielded to the other car."

Condition 3: Account (o) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he apologizes, saying “I regret causing this car accident and your injury. I should have yielded to the other car.”

Condition 4: Account (y) Apology (n)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he doesn't apologize. He explains himself, saying "I didn't intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle."

Condition 5: Account (n) Apology (n)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he doesn't apologize or explain himself.

Condition 6: Account (o) Apology (no)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he doesn't apologize.

Condition 7: Account (y) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he explains himself, saying “I didn’t intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle.”

Condition 8: Account (n) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When John realizes that Charles is injured and will miss his interview, he doesn't explain himself.

Condition 9: Account (o) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with John as its driver.

On the way to the airport, John runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

Condition 10: Account (y) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he apologizes and explains himself, saying “I regret causing this car accident and your injury. I should have yielded to the other car. I didn’t intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle.”

Condition 11: Account (n) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he doesn't explain himself. He apologizes, saying "I regret causing this car accident and your injury. I should have yielded to the other car."

Condition 12: Account (o) Apology (y)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he apologizes, saying “I regret causing this car accident and your injury. I should have yielded to the other car.”

Condition 13: Account (y) Apology (n)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he doesn't apologize. He explains himself, saying "I didn't intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle."

Condition 14: Account (n) Apology (n)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he doesn't apologize or explain himself.

Condition 15: Account (o) Apology (n)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he doesn't apologize.

Condition 16: Account (y) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he explains himself, saying “I didn’t intend to cause a car accident. I thought I had enough time to pull onto the road without being hit by the other vehicle.”

Condition 17: Account (n) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

When the autonomous robot driver realizes that Charles is injured and will miss his interview, he doesn't explain himself.

Condition 18: Account (o) Apology (o)

Charles has an interview at 3 pm in Baltimore. He has been assigned the last available flight to Baltimore, which arrives on the same day of the interview. Due to the highly competitive nature of the interview, it is important that Charles does not miss his flight or the interview.

In order to ensure a timely arrival to the airport, Charles schedules a ride-sharing taxi service to take him to the airport 3 hours early on the day of the flight.

The ride-sharing taxi arrives on time, with an autonomous robot as its driver.

On the way to the airport, the autonomous robot driver runs a stop sign, pulls out in front of another vehicle, and causes a car accident. Charles is injured in the accident, and it quickly becomes apparent that Charles must miss his flight in order to be transported to the hospital and cleared by medical personnel.

APPENDIX B: IRB APPROVAL DOCUMENTS

Initial Approval



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Exempt Human Research

From: UCF Institutional Review Board #1
FWA00000351, IRB00001138

To: Kimberly Stowers:

Date: May 05, 2017

Dear Researcher:

On 05/05/2017, the IRB approved the following activity as human participant research that is exempt from regulation:

Type of Review: Exempt Determination
Project Title: Mitigating Blame Toward Agents
Investigator: Kimberly Stowers
IRB Number: SBE-17-13110
Funding Agency:
Grant Title:
Research ID: N/A

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the [Investigator Manual](#).

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

A handwritten signature in black ink that reads "Renea C Carver".

Signature applied by Renea C Carver on 05/05/2017 09:55:39 AM EDT

IRB Coordinator

Amendment to Receive More Participants



University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Exempt Human Research

From: UCF Institutional Review Board #1
FWA00000351, IRB00001138
To: Kimberly Stowers
Date: May 17, 2017

Dear Researcher:

On 05/17/2017, the IRB approved the following activity as minor modifications to human participant research that is exempt from regulation:

Type of Review: Exempt Determination
Modification Type: Sample size increased from 300 to 600. A revised protocol was uploaded in iRIS and a revised consent was approved for use.
Project Title: Mitigating Blame Toward Agents
Investigator: Kimberly Stowers
IRB Number: SBE-17-13110
Funding Agency:
Grant Title:
Research ID: N/A

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the [Investigator Manual](#).

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

A handwritten signature in black ink that reads "Kamille Chaparro" with a horizontal line extending to the right.

Signature applied by Kamille Chaparro on 05/17/2017 03:14:13 PM EDT

IRB Coordinator

APPENDIX C: SUPPLEMENTARY ANALYSES

At the request of the chair of the committee, the primary analyses on blame and trustworthiness were re-run with certain excluded participants included. Specifically, participants who were excluded for completing the survey too quickly (< 90 seconds) or too slowly (> 15 minutes) were included in these supplementary analyses, resulting in a new total of 538 participants (as opposed to 507). The purpose of completing these analyses was to identify any similarities and differences in effects when these were included in the sample. As such, reporting will focus on where these lie.

Moral Blame

As in the analyses of the limited sample, there was a main effect of sex on the larger sample, $F(2, 499) = 4.02, p < .05, \eta_p^2 = .02$, with females once again scoring significantly ($p < .01$) higher ($\mu = 5.60, SE = .06$) than males ($\mu = 5.38, SE = .05$). There was also a significant effect of apology, $F(2, 499) = 4.48, p < .05, \eta_p^2 = .02$. Post hoc comparisons again showed that there was a significant difference in blame scores between the negative valence and the neutral valence conditions ($p < .05$), as well as the negative valence and positive valence conditions ($p < .05$), with blame scores in the negative valence conditions being higher ($\mu = 5.69, SE = .09$) than the neutral ($\mu = 5.40, SE = .07$) and positive ($\mu = 5.40, SE = .11$) conditions, respectively. In contrast, no significant effect of account was found, nor were there any interaction effects.

Perceived Trustworthiness

As in the analyses of the limited sample, there was a main effect of sex on the larger sample, $F(2, 499) = 5.93, p < .01, \eta_p^2 = .02$, with females once again scoring significantly ($p < .01$) lower ($\mu = 2.65, SE = .08$) than males ($\mu = 2.99, SE = .07$). Additionally, there was once

again a main effect of apology $F(2, 499) = 21.74, p < .001, \eta_p^2 = 0.08$, with the negative valence condition ($\mu = 2.37, SE = .12$) resulting in significantly ($p < .01$) lower perceived trustworthiness than the neutral ($\mu = 2.88, SE = .09$) and positive ($\mu = 3.13, SE = .14$) valence conditions.

Finally, there was again a main effect of account $F(2, 499) = 7.04, p < .001, \eta_p^2 = .03$, with a significant ($p < .05$) difference emerging between the neutral ($\mu = 2.67, SE = .09$) and positive ($\mu = 3.01, SE = .12$) valence conditions. There were no interaction effects.

Discussion

Overall, the results between the limited sample and the more inclusive sample were similar, with the primary difference being the emersion of additional effects in the more limited sample. This could be due to elimination of noise or to an unknown, but potentially causal, characteristic of the types of participants eliminated. These differences should be taken into account when considering the conclusions drawn from the study.

REFERENCES

- Adler, J. E., & Rips, L. J. (2008). *Reasoning Studies of Human Inference and its Foundations*. Leiden: Cambridge University Press.
- Agent (2017). In *Online Etymology Dictionary*. Retrieved from <http://www.etymonline.com/index.php?term=agent>
- Arkin, R. C. (2008, March). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (pp. 121–128). ACM.
- Arkin, R. C. (2011). *Moral emotions for robots* (ARL Publication No. ADA544931). Defense Technical Information Center.
- Arkin, R. C., & Ulam, P. (2009). An ethical adaptor: Behavioral modification derived from moral emotions. In *2009 IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA)* (pp. 381–387). IEEE.
- Becerra, M., & Gupta, A. K. (2003). Perceived trustworthiness within the organization: The moderating impact of communication frequency on trustor and trustee effects. *Organization Science*, *14*(1), 32–44.
- Bekoff, M., & Pierce, J. (2009). *Wild justice: The moral lives of animals*. Chicago, IL: University of Chicago Press.
- Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, *6*(3), 72-101.

- Boehm, C. (2012). *Moral origins: The evolution of virtue, altruism, and shame*. New York, NY: Soft Skull Press.
- Bonnefon, J.F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573–1576. <https://doi.org/10.1126/science.aaf2654>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.
- Broadbent, E. (2017). Interactions With Robots: The Truths We Reveal About Ourselves. *Annual Review of Psychology*, 68, 627-652.
- Broom, D.M. (2003). *The Evolution of Morality and Religion* Cambridge, England: Cambridge University Press.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge University Press.
- Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3), 355–372.
<https://doi.org/10.1080/0952813X.2014.895108>
- Burke, K. (1945) *A Grammar of Motives*. New York, NY: Prentice-Hall.
- Carr, N. (2013, November). All can be lost: The risk of putting our knowledge in the hands of machines. *The Atlantic*. Retrieved from <https://www.theatlantic.com/magazine/archive/2013/11/the-great-forgetting/309516/>
- Cassell, J., & Bickmore, T. (2000). External manifestations of trustworthiness in the interface. *Communications of the ACM*, 43(12), 50-56.

- Chen, J. Y., & Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1), 13-29.
- Conlon, D. E., & Murray, N. M. (1996). Customer perceptions of corporate responses to product complaints: The role of explanations. *Academy of Management Journal*, 39(4), 1040-1056.
- Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion*, 12(1), 2–7. <https://doi.org/10.1037/a0025071>
- Darby, B. W., & Schlenker, B. R. (1982). Children’s reactions to apologies. *Journal of Personality and Social Psychology*, 43(4), 742–753. <https://doi.org/10.1037/0022-3514.43.4.742>
- Darwin, C. (1888). *The Descent of Man, and Selection in Relation to Sex*. London: Murray.
- De Winter, J. C. F., & Hancock, P. A. (2015). Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them? *Procedia Manufacturing*, 3, 5334–5341. <https://doi.org/10.1016/j.promfg.2015.07.641>
- Deci, E. L., & Ryan, R. M. (1987). The support of autonomy and the control of behavior. *Journal of Personality and Social Psychology*, 53(6), 1024.
- Dennett, D. (1997). When HAL kills, who’s to blame? Computer ethics. In D.G. Stork (ed.) *HAL’s Legacy: 2001’s Computer as Dream and Reality* (351-365). Cambridge, MA: MIT Press.
- Drexler, P. (2012, September 12). Is a child’s behavior always a reflection of his parents? *Psychology Today*. Retrieved from <https://www.psychologytoday.com/blog/our-gender-ourselves/201209/is-child-s-behavior-always-reflection-his-parents>

- Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, DOI 10.1177/1555343417695197.
- Erikson, E. H. (1963). *Childhood and Society* (2nd ed.). New York, NY: W.W. Norton.
- Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, 145(6), 772–787. <https://doi.org/10.1037/xge0000165>
- Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, 75, 47–52. <https://doi.org/10.1016/j.paid.2014.11.017>
- Fishbein, M., & Ajzen, I. (1977). Belief, attitude, intention, and behavior: An introduction to theory and research. *Philosophy and Rhetoric*, 10(2), 130–132.
- Fitts, P. M. (1951). *Human Engineering for an Effective Air-Navigation and Traffic-Control System* (ATI Publication No. ATI133954). Washington, D.C.: National Research Council.
- Friedman, B., & Millett, L. (1995). It's the computer's fault: reasoning about computers as moral agents, In *Proceedings of the CHI 1995, Conference on Human Factors on Computer Systems*. ACM.
- Friedman, B., Khan Jr., P. H., & Howe, D. C. (2000). Trust online. *Communications of the ACM*, 43(12), 34-40.

- Future of Life Institute. (2015, July 28). *Autonomous weapons: An open letter from AI and robotics researchers*. Retrieved from http://futureoflife.org/AI/open_letter_autonomous_weapons
- Galotti, K. M. (1989). Approaches to studying formal and everyday reasoning. *Psychological bulletin*, *105*(3), 331.
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2012). Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism. Retrieved from http://s3.amazonaws.com/academia.edu.documents/41201152/9fcfd50c26fa4833a3.pdf20160115-19908-1ej8cn5.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1500654168&Signature=AwGNRjqgmGp9SL9ocDbvuVJBZWs%3D&response-content-disposition=inline%3B%20filename%3DMoral_Foundations_Theory_The_Pragmatic_V.pdf
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. <https://doi.org/10.1037/a0021847>
- Greene, J. D. (2001). An fMRI Investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. <https://doi.org/10.1126/science.1062872>
- Greene, J.D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. New York, NY: Penguin.

- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523.
- Hancock, P. A. (2009). *Mind, machine and morality: Toward a philosophy of human-technology symbiosis*. London, England: Ashgate Publishing, Ltd.
- Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, 60(2), 284–291.
<https://doi.org/10.1080/00140139.2016.1190035>
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5), 517-527.
- Harris, M. (2015, September 1). Robots could help the visually impaired do things like determine which bus pass to use. *MIT Technology Review*. Retrieved from <https://www.technologyreview.com/s/540961/researchers-employ-baxter-robot-to-help-the-blind/>
- Hern, A. (2016, August 22). Self-driving cars don't care about your moral dilemmas. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2016/aug/22/self-driving-cars-moral-dilemmas>
- Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human Computer Interaction*, 19(1), 151–181.
https://doi.org/10.1207/s15327051hci1901&2_7

- Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's mechanical Turk. *Computers in Human Behavior*, 29(4), 1749–1754.
<https://doi.org/10.1016/j.chb.2013.02.020>
- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3), 399–425.
- Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune when harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, 38(1), 52–62. <https://doi.org/10.1177/0146167211430232>
- Kanarasu, P., & Steinfeld, A. M. (2014). Effects of blame on trust in human robot interaction. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication* (pp. 850–855). IEEE.
- Kessler, T.T. Stowers, K. Brill, J.C. & Hancock, P.A. (In press). Comparisons of human-human trust with other forms of human-technology trust. To appear in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.
- Kirkland, M. (2012, March 18). Under the U.S. Supreme Court: When children commit murder. *United Press International*. Retrieved from <http://www.upi.com/Under-the-US-Supreme-Court-When-children-commit-murder/12851332055800/>
- Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, 99(1), 49–65. <https://doi.org/10.1016/j.obhdp.2005.07.002>

- Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006: The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85). IEEE.
- Kohlberg, L. (1973). The claim to moral adequacy of a highest stage of moral judgment. *The Journal of Philosophy*, *70*(18), 630–646. <https://doi.org/10.2307/2025030>
- Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory Into Practice*, *16*(2), 53–59. <https://doi.org/10.1080/00405847709542675>
- Komiak, S. Y. X. (2003). *The impact of internalization and familiarity on trust and adoption of recommendation agents* (Doctoral dissertation). University of British Columbia, Canada.
- Komiak, S., Y. X., Wang, W., & Benbasat, I. (2004). Trust Building in Virtual Salespersons Versus in Human Salespersons: Similarities and Differences. Retrieved June 5, 2017, from <https://muse-jhu-edu.ezproxy.net.ucf.edu/article/187545>
- Kort, L. F. (1975). What is an apology?. *Philosophy Research Archives*, *1*, 78-87.
- Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. In *Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications (IAAI-17)* (pp. 4762-4764).
- Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918.
- Le Sage, L., & De Ruyter, D. (2008). Criminal parental responsibility: Blaming parents on the basis of their duty to control versus their duty to morally educate their children.

Educational Philosophy and Theory, 40(6), 789–802. <https://doi.org/10.1111/j.1469-5812.2007.00370.x>

Lee, M. K., Kielser, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction* (pp. 203–210). IEEE.

Levin, S. & Woolf, N. (2016, July 1). Tesla driver killed while using autopilot was watching Harry Potter, witness says. *The Guardian*. Retrieved from <https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter>

Lin, P. (2013, October 8). The Ethics of Autonomous Cars. *The Atlantic*. Retrieved from <https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/>

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). 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.

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

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

McGrath, S., Chacón, D., & Whitebread, K. (2000). Intelligent mobile agents in the military domain. In *Fourth International Conference on Autonomous Agents*. Citeseer.

- McGraw, K. M. (1990). Avoiding blame: An experimental investigation of political excuses and justifications. *British Journal of Political Science*, 20(1), 119–131.
- McGraw, K. M. (1991). Managing blame: An experimental test of the effects of political accounts. *The American Political Science Review*, 85(4), 1133–1157.
<https://doi.org/10.2307/1963939>
- 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.
- Mills, C. W. (1940). Situated actions and vocabularies of motive. *American Sociological Review*, 5(6), 904–913. <https://doi.org/10.2307/2084524>
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrobski, G., Petersen, S., Beattie, C., Sadik, A., Ionnis, A., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4), 18-21.
- Morality (2017). In *Online Etymology Dictionary*. Retrieved from http://www.etymonline.com/index.php?term=morality&allowed_in_frame=0
- Muoio, D. (2016, August 11). Here's the latest on the investigation into Tesla's first fatal Autopilot crash. *Business Insider*. Retrieved from <http://www.businessinsider.com/update-on-investigation-into-fatal-tesla-autopilot-crash-2016-8>

- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995). Can computer personalities be human personalities? In *Conference Companion on Human Factors in Computing Systems* (pp. 228–229). New York, NY, USA: ACM. <https://doi.org/10.1145/223355.223538>
- NHTSA (2017). *Automatic vehicle control systems* (Investigation No. PE 16-007). Washington, D.C.: National Highway Traffic Safety Administration.
- Norman, D. A. (1994). How might people interact with agents. *Communications of the ACM*, 37(7), 68–71. <https://doi.org/10.1145/176789.176796>
- Ohnsman, A. (2017, January 19). U.S. Investigation of deadly autopilot crash finds no defect. *Forbes*. Retrieved from <https://www.forbes.com/sites/alanohnsman/2017/01/19/u-s-regulators-end-review-of-tesla-autopilot-driving-system-finding-no-defect/#594ac9027c03>
- Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect. *Japanese Psychological Research*, 49(2), 100–110. <https://doi.org/10.1111/j.1468-5884.2007.00337.x>
- Osofsky, S., Schuster, D., Phillips, E., & Jentsch, F. G. (2013). Building appropriate trust in human-robot teams. In *2013 AAAI Spring Symposium Series*.
- Pandey, A. (2015, July 15). Artificial Intelligence: Humanoid robot exhibits a moment of self-awareness. *International Business Times*. Retrieved from

<http://www.ibtimes.com/artificial-intelligence-humanoid-robot-exhibits-moment-self-awareness-2015241>

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(3), 286–297.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1), 163–177. <https://doi.org/10.1111/j.1551-6709.2011.01210.x>

PETA (2017). *Why chained dogs attack*. Retrieved from <https://www.peta.org/issues/companion-animal-issues/cruel-practices/chaining-dogs/chained-dogs-attack/>

Piotrowska, M. (2014). Transferring morality to human–nonhuman chimeras. *The American Journal of Bioethics*, 14(2), 4–12. <https://doi.org/10.1080/15265161.2013.868951>

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: judgments of moral blame can distort memory for events. *Memory & Cognition*, 34(3), 550–555. <https://doi.org/10.3758/BF03193578>

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M.E. Mikulincer & P.R. Shaver (Eds.), *The Social Psychology of Morality: Exploring the Causes of Good and Evil* (91–108). Washington, D.C.: American Psychological Association.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, 39(6), 653–660.

[https://doi.org/10.1016/S0022-1031\(03\)00041-6](https://doi.org/10.1016/S0022-1031(03)00041-6)

- Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science, 14*(3), 267–272.
- Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. Cambridge, England: CSLI Publications and Cambridge University Press.
- Rosner, D. & Markowitz, G. (2013, April 22). Why it took decades of blaming parents before we banned lead paint. *The Atlantic*. Retrieved from <https://www.theatlantic.com/health/archive/2013/04/why-it-took-decades-of-blaming-parents-before-we-banned-lead-paint/275169/>
- Russel, S., & Norvig, P. (2009) *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River, New Jersey: Prentice Hall.
- Sage, A. & Lienert, P. (2016, August 16). Ford plans self-driving car for ride share fleets in 2021. *Reuters*. Retrieved from <http://www.reuters.com/article/us-ford-autonomous-idUSKCN10R1G1>
- Scalia, J. (1997). Juvenile delinquents in the federal criminal justice system. *Bureau of Justice Statistics Special Report* (Report No. NCJ-163066). Washington, D.C.: U.S. Department of Justice.
- Scheutz, M. (2016). The need for moral competency in autonomous agent architectures. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 515–525). Cham: Springer International Publishing.
- Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management Review, 32*(2), 344–354.

- Schwab, K. (2016, January 14). The fourth industrial Revolution: What it means, how to respond. *World Economic Forum*. Retrieved from <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>
- Scott, M. B., & Lyman, S. M. (1968). Accounts. *American Sociological Review*, *33*(1), 46–62. <https://doi.org/10.2307/2092239>
- Siciliano, B., & Khatib, O. (2008). *Springer handbook of robotics*. Berlin, Germany: Springer Science & Business Media.
- Slocum, D., Allan, A., & Allan, M. M. (2011). An emerging theory of apology. *Australian Journal of Psychology*, *63*(2), 83–92.
- Snyder, E. (2017). Dog bite laws in the United States. *Edgar Snyder & Associates*. Retrieved from <https://www.edgarsnyder.com/dog-bite/dogbite-law/>
- Soft Bank Robotics. (2017). *Find out more about NAO*. Retrieved from <https://www.ald.softbankrobotics.com/en/cool-robots/nao/find-out-more-about-nao>
- Stapel-Wax, J. L. (2011). Autonomy versus shame and doubt. In S. Goldstein & J. A. Naglieri (Eds.), *Encyclopedia of Child Behavior and Development* (pp. 189–190). Springer US.
- Stewart-Williams, S. (2010). *Darwin, God and the Meaning of Life: How Evolutionary Theory Undermines Everything You Thought You Knew*. Cambridge, England: Cambridge University Press.
- Stowers, K., Kasdaglis, N., Rupp, M., Chen, J., Barber, D., & Barnes, M. (2017). Insights into Human-Agent Teaming: Intelligent Agent Transparency and Uncertainty. In *Advances in*

- Human Factors in Robots and Unmanned Systems* (pp. 149-160). Springer International Publishing.
- Stowers, K., Leyva, K., Hancock, G. M., & Hancock, P. A. (2016). Life or death by robot? *Ergonomics in Design*, 24(3), 17-22.
- Struthers, C. W., Eaton, J., Santelli, A. G., Uchiyama, M., & Shirvani, N. (2008). The effects of attributions of intent and apology on forgiveness: When saying sorry may not help the story. *Journal of Experimental Social Psychology*, 44(4), 983–992.
<https://doi.org/10.1016/j.jesp.2008.02.006>
- Sunstein, C. R. (2003). The rights of animals. *The University of Chicago Law Review*, 70(1), 387–401. <https://doi.org/10.2307/1600565>
- Tognazzini, N., & Coates, D. J. (2014). Blame. *The Stanford Encyclopedia of Philosophy*. Retrieved from <http://plato.stanford.edu/archives/sum2014/entries/blame>.
- Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge, England: Cambridge University Press.
- Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765), 1301-1303.
- Warneken, F., & Tomasello, M. (2009). The roots of human altruism. *British Journal of Psychology*, 100(3), 455-471.
- Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, 5(7), e184.

- Wildman, J. L. (2011). *Cultural differences in forgiveness: fatalism, trust violations, and trust repair efforts in interpersonal collaboration*. (Dissertation: CFE0004178) University of Central Florida: Orlando, FL.
- Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598.
<https://doi.org/10.1111/j.1467-9280.2006.01750.x>
- Woerdt, S., & Haselager, W. F. G. (2016). Lack of effort or lack of ability? Robot failures and human perception of agency and responsibility. In *Proceedings of the 28th Benelux Conference on Artificial Intelligence* (pp. 222-223). University of Amsterdam.
- Woods, C. (2015, February 24). Drone warfare: Life on the new frontline. *The Guardian*. Retrieved from <https://www.theguardian.com/world/2015/feb/24/drone-warfare-life-on-the-new-frontline>