MORAL MACHINE
Perception of Moral Judgment Made by Machines

EDMOND AWAD

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
JUL 31 2017
LIBRARIES
ARCHIVES

B.S. in Informatics Engineering, Tishreen University, 2007
M.Sc. in Computing and Information Science, Masdar Institute, 2011
Ph.D. in Computing and Information Science, Masdar Institute, 2015

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology.

JUNE 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

Signature of Author: Signature redacted
Program in Media Arts and Sciences
May 12, 2017

Certified by: Signature redacted
Iyad Rahwan
Associate Professor of Media Arts and Sciences
Thesis Supervisor

Signature redacted
Accepted by:
Pattie Maes
Professor of Media Arts and Sciences
Academic Head, Program in Media Arts and Sciences

MORAL MACHINE
Perception of Moral Judgment Made by Machines

EDMOND AWAD

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 12, 2017 in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology.

ABSTRACT

While technological development of vehicular autonomy has been progressing rapidly, a parallel discussion has emerged with regard to the moral implications of a future wherein people hand over to autonomous machines the controls to a mode of transportation. These discussions have entered a new phase with the U.S. Department of Transportation (DoT) releasing a 15-point policy that requires manufacturers to explain how their AVs will handle "ethical considerations". However, there is a huge gap in our understanding of the ethical perception of AI, as there have been few large-scale empirical studies on human moral perception of outcomes to autonomous vehicle moral dilemmas. Additionally, public engagement is a very important piece of the puzzle, especially given the emotional salience of traffic accidents. With that in mind, I co-developed the "Moral Machine" (`http://moralmachine.mit.edu`). *Moral Machine* is a platform for gathering a human perspective on moral decisions made by machine intelligence, such as AVs. The web site went viral, and got covered in various media outlets. This web site has also been a valuable data collection tool, allowing us to collect the largest dataset on AI ethics ever collected in history (with 30 million decisions by over 3 million visitors, so far). This thesis will introduce the *Moral Machine* platform as a data-gathering platform. Moreover, insights about the human perception of the different routes to full automation will be covered in the thesis, with the data collected through other online platforms.

Thesis Supervisor: Iyad Rahwan
Title: Associate Professor of Media Arts and Sciences

MORAL MACHINE
Perception of Moral Judgment Made by Machines

EDMOND AWAD

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 12, 2017 in partial fulfillment of the requirements for the degree of Master of Science in Media Arts and Sciences at the Massachusetts Institute of Technology.

The following people served as readers for this thesis:

Thesis Reader: Signature redacted
Joshua Greene
Professor of Psychology
Harvard University

Thesis Reader: Signature redacted
Joshua B. Tenenbaum
Professor of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Dedicated to the machines of the future.
May you find your moral compass.

ACKNOWLEDGMENTS

First and foremost, I would like to express my special gratitude and appreciation to my adviser Dr. Iyad Rahwan. It has been an honor to be his first graduating student at MIT. His guidance, motivation, and support has extended beyond the content of this thesis.

I would also like to thank my readers, Prof. Joshua Green, and Prof. Joshua B. Tenenbaum for their invaluable comments on my thesis, and their positive support of my research work. I am sincerely honored to have the approval of such two established researchers on my Master's thesis.

The work in my thesis is the result of various fruitful collaborations. Special thanks to Prof. Jean-François Bonnefon and Dr. Azim Shariff. Their contributions were significant to the design, the analysis, and the write-up of the study in Chapter 3 and the design of Moral Machine. I would like also to thank my lab mates Sohan Dsouza and Pai-Ju Chang. Sohan co-designed the pilot experiments described in Chapter 3. Sohan was also the main co-designer and co-developer of the Moral Machine web site. We jointly co-authored the descriptions of Moral Machine in Section 4.2 as part of a joint paper planned for submission. Pai-Ju made a significant contribution on the development of Moral Machine and helped prettify the result figures in Chapter 4. I would like to thank Max Kleiman-Weiner who has been a great collaborator. He co-designed the experiments, and we jointly co-authored the study along with Bonnefon and Azim in Chapter 3.

The final version of this thesis has benefited to a great extent from the comments and the feedback of very close friends who generously spent time reading one (or more) chapters of an earlier version of this thesis. Thank you Sophie Chu, Karen Scott, Juliana Nazaré, Martin Saveski, Ilse Verdiesen, Maggie Church, Eric Chu, and Cristian Ignacio Jara Figueroa.

At MIT Media Lab, I would like to thank the members and the visitors of the Scalable Cooperation group (former and current): Pinar Yanardag, Andres Abeliuk, Lijun Sun, Nick Obradovich, Sydney Levine, Manuel Cebrian, Morgan Frank, Richard Kim, Neil Gaikwad, Bjarke Felbo, Antonio Fernández Anta, Hye-Jin Youn, Lorenzo Coviello, and Yves-Alexandre de Montjoye. Each one of you has contributed to an enjoyable and constructive experience during these two years. Special thanks to Amna Carriero who has always been caring and loving.

Outside the Scalable Cooperation group, there have been many Media Lab students who were close friends and have contributed to my pleasant Master's time as well. Thank you Mina Soltangeis, Aamena Alshamsi, Anneli, Rane, Natasha Jaques, Abdullah Almaatouq, Abdulrahman Alotaibi, Dhaval Adjodah, Alejandro Noriega, Carmelo Presicce, and Caroline Jaffe. Thank you Keira Horowitz and Linda Peterson for the continuous help.

Outside Media Lab, there have been many people that made my stay in Boston unforgettable and joyful. Thank you Shaheen Tomah, Mohamed Qasim, Shahd Labib, Mohamed Radwan, Cynthia Hajal, Rana Elkahwagy, Nour Elraghy, Danielle Atwell, Majd Jradeh, and Amira Awad.

Many people who are outside Boston, have also been very supporting and loving. Thanks to Nabil Kenan, Fadi Awad, Joe Hermez, Louis Louis, Erisa Karafili, Daniel Naaman, Rani Basna, Hiba Nassar, Hussam Awad, Eliane Naaman, Wael Awad, Iyad Nasrah, Adon Naaman, Basel Moussa, Abdullah Nasrah, Marin Nassar, Maha Nassar, Nadim Nassar, Rawan Awad, Feras Homaysi, Loai Anttar, Bedoor Al Shebli, Abdo Maisari, Eid Hatem, Yona Haddad, Alice Hatem, Tarek Daher, Saji Alhaddad, Manhal Abboud, Fuad Toumeh, Maya Naaman, Nermin Toumeh, and Firas Awad.

Finally, my greatest gratitude goes to my family for their continuous love, encouragement and support. They have cherished with me every great moment and provided support whenever I needed it. For my parents, Adel Awad and Istiklal Hatem who raised me with a love of science and supported me in all my pursuits. For my sister, Maya and my brother, Nadi for being such loving and supportive siblings. Special thanks to my brother-in-law, Mario and my beautiful niece Eliana, who joined the family during my Master's. I was fortunate to spend time with both during my program. I also thank my grandmother, Alice for her sincere and consistent prayers and love. I also thank my late grandmother Najiba, who has always been unconditionally caring and loving. Her memory will be with me always.

Edmond Awad
Cambridge, MA, US
May 12th, 2017

CONTENTS

1 INTRODUCTION 17
  1.1 Research Question 19
  1.2 Thesis Overview 20
2 BACKGROUND 21
  2.1 Normative Ethics 22
    2.1.1 Kantian Deontology 23
    2.1.2 Utilitarianism as the Consequentialism Paradigm 24
  2.2 Trolley Problem 24
  2.3 Descriptive Ethics 26
  2.4 Machine Ethics 27
  2.5 Towards Descriptive Machine Ethics 29
3 MORAL RESPONSIBILITY ACROSS LEVELS OF VEHICLE AUTOMATION 31
  3.1 Results 34
    3.1.1 Relative allocation of responsibility and blame 34
    3.1.2 Absolute allocation of responsibility and blame 36
  3.2 Discussion 37
  3.3 Methods 38
4 MORAL MACHINE AS A MASSIVE ONLINE EXPERIMENTATION (MOE) TOOL 41
  4.1 Massive Online Experiments (MOE) 42
  4.2 Moral Machine 42
    4.2.1 Requirements 43
    4.2.2 Implementation 43
  4.3 Preliminary results 52
5 DISCUSSION AND FUTURE WORK 61
  5.1 Limitations of the Trolley Problem as a Paradigm to Study Machine Ethics 61
  5.2 Contributions 63
  5.3 Future Work 65
A APPENDIX: SUPPLEMENTAL INFORMATION FOR AUTOMATION REGIMES STUDY 69
  A.1 Covariates Balance 69
  A.2 Order Balance 71
  A.3 Full Results 73
  A.4 Vignettes 75

BIBLIOGRAPHY 79

LIST OF FIGURES

Figure 1 A row of Google self-driving cars [1]. 17

Figure 2 An outline of ethics as a subfield of philosophy. This chapter provides a background for parts in red. This figure is not meant to provide a comprehensive outline, but an illustrative one to show how different parts relate. That said, *Descriptive Ethics* is considered by some as not part of philosophy of ethics. Further, the dashed box and links (for *Descriptive Machine Ethics*) are not part of the outline, but added here to illustrate how the main topic of this thesis relates to the rest. 22

Figure 3 Jeremy Bentham and Immanuel Kant, leading figures of *consequentialism* and *deontology*, respectively. 23

Figure 4 A visual depiction of the *Trolley* Problem – *Moral Machine* interface. (a) The original and most famous version – Spur. (b) A famous variant of the Trolley Problem – Footbridge. (c) Another famous variant of the Trolley Problem – Loop. 25

Figure 5 Four levels of automation for autonomous vehicles: Regular car (RC), Guardian Angel (GA), Autopilot (AP), and Fully Autonomous (FA) regimes. These levels broadly correspond to the classifications of the NHTSA and SAE International [79]. Levels 1 (RC) and 4 (FA) involve only one driver and no standby. Levels 2 (GA) and 3 (AP) involve two agents: the main driver (shown in red), and the standby driver (shown in blue). The causal structure of the decision making is different across the four levels, but in each case, the main driver makes a decision (in this article, stay or swerve), and the standby driver (if any) then decides whether to override the decision of the main driver. 32

Figure 6 95% confidence intervals of the difference in ratings of causal responsibility and blameworthiness between the Industry and the User, across four automation regimes, after a suboptimal action or inaction. 35

Figure 7 Causal responsibility and blame-worthiness ratings of user, car, programmer, and company. Values for programmer in regular cars (RC) and fully-autonomous (FA) were not collected. The user receives highest ratings under the Autopilot (AP) regime, while the company receives highest ratings under the Guardian Angel (GA) regime. 36
Figure 8 *Moral Machine* interface. 44
Figure 9 The number of users of Moral Machine from each (a) country, and from each (b) US state. 52
Figure 10 Overview of the demographics of *Moral Machine* users. 52
Figure 11 Distributions of the "stated" preferences of *Moral Machine* users over the nine dimensions. 54
Figure 12 The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by gender. 55
Figure 13 The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by political views. 56
Figure 14 The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by religious views. 57
Figure 15 World map and US map highlighting (a)-(b) utilitarianism – saving more people, (c)-(d) tendency to saving passengers as opposed to pedestrians, (e)-(f) non-interventionism – tendency to leave the AV on its track and avoid swerving, and (g)-(h) saving the lawful pedestrians. 58
Figure 16 Vignette for Guardian Angel with Suboptimal Inaction. 75
Figure 17 Vignette for Guardian Angel with Suboptimal Action. 75
Figure 18 Vignette for Autopilot with Suboptimal Inaction. 76
Figure 19 Vignette for Autopilot with Suboptimal Action. 76
Figure 20 Vignette for Regular Car with Suboptimal Inaction. 77
Figure 21 Vignette for Regular Car with Suboptimal Action. 77
Figure 22 Vignette for Fully Autonomous car with Suboptimal Inaction. 77

Figure 23 Vignette for Fully Autonomous car with Suboptimal Action. 78
Figure 24 Questions asked for the conditions where Agent is the *Robocar*. Other cases replaced Robocar with Car (when the scenarios is about a regular car, Robocar/car company, or Robocar/car programmer.) 78

LIST OF TABLES

Table 1 Pros/Cons of the use of Trolley Problem as a paradigm 64
Table 2 OLS Results: Propensity for Treatment Given Covariates 70
Table 3 Order Balance – AP and GA – Car 71
Table 4 Order Balance – AP and GA – Company 71
Table 5 Order Balance – AP and GA – Programmer 72
Table 6 Order Balance – FA and RC – Car 72
Table 7 Order Balance – FA and RC – Company 72
Table 8 Full Results – Relative Allocation (Industry - User) of Causal Responsibility and Blame. The mean of per-participant difference between Industry and User attribution and the 95% confidence intervals of the differences (between parenthesis) are shown for each of the four automation paradigms: regular car (RC), Guardian Angel (GA), Autopilot (AP), and fully autonomous car (FA). Results are aggregated over Suboptimal Inaction (omission leading to death of five people) and Suboptimal Action (commission leading to death of five people). 73
Table 9 Full Results – Absolute Allocation of Causal Responsibility and Blame. The mean and the 95% confidence intervals of the means (between parenthesis) are shown for each of the four automation paradigms: regular car (RC), Guardian Angel (GA), Autopilot (AP), and fully autonomous car (FA). Results are aggregated over each of the four agents: User, Car, Company, and Programmer. 74

1

INTRODUCTION

> "What we really need is a sound way to teach our machines to be ethical.
> The trouble is that we have almost no idea how to do that."
>
> –*Gary Marcus, The New Yorker*

Figure 1: A row of Google self-driving cars [1].

Robots and other Artificial Intelligence (AI) systems are in the process of transitioning from performing well-defined tasks in closed environments, to becoming significant physical actors in the real world. No longer confined within the walls of factories, robots will permeate the urban environment, moving people and goods around, and performing all kinds of tasks alongside humans. Nothing exemplifies this transition more than the imminent rise of Autonomous Vehicles (AVs).

AVs promise numerous social and economic advantages. They are expected to vastly increase the efficiency of our transportation infrastructure, reducing pollution, and freeing up millions of man-hours of productivity. They also promise to drastically cut rates of death and injury from traffic accidents [25, 82].

With millions of AVs potentially moving around our urban environment, AVs are arguably the first human-made artifacts to make autonomous decisions with potential life-and-death consequences on a broad scale. This marks a qualitative shift in the consequences of design choices made by engineers.

In particular, AVs will generate negative externalities – consequences affecting third parties not involved in their adoption. An example of these externalities is an AV that prioritizes saving the lives of its passengers over pedestrians. Negative externalities can greatly influence economic growth and social life. While an obvious solution to limit those externalities is to employ policing by an authority, machine policing will be a tricky task for various reasons.

For one thing, machines are still seen as *blackboxes*: it is unclear how they process their input and how they make decisions, sometimes even to those who programmed the machines [14]. This makes it difficult to implement precautionary procedures to limit potential externalities, and it prompts other concerns like accountability. Being accustomed to their old role, we are not yet fully used to attributing moral agency to machines, and thus we still have a psychological barrier that prevents us from holding them accountable. The non-transparency of machines would make it even harder to accept holding them responsible, even within the new role they play. Further, machines will be in a constant process of learning, which makes it unclear whether their actions at any time reflect their long-term behavior. Thus, prior testing and certification might not be sufficient. Another concern is that AI systems have been shown to be biased when making decisions [73], and it is unclear whether it is due to their design or due to learning from human biases. These concerns are not only valid but are also alarming, and have strong potential to slow down the transition of AI into the role of the primary agent, unless the public starts seeing them getting addressed satisfactorily soon.

Unfortunately, there are some challenges that stand in the way of addressing these concerns. These challenges can be broadly summed up by a huge gap between humanities from one side and engineering from the other side. While ethicists, legal scholars, and moral philosophers are capable of diagnosing moral hazards and identifying violations of laws and norms, they are not used to framing their expectations in a programmable way. On the other hand, engineers are not always capable of communicating the behavior of their systems using the same language that the ethicists and legal theorists use and understand. Furthermore, it is likely that spelling out the normative ideals of their systems using the ethicists' language would expose incoherence in these ideals.

This gap contributes to the lack of a comprehensive moral code for machines. It is unclear what values, principles, and ideals we want for these machines to have when they start making decisions on their own. The gap also contributes to the lack of mechanism to embed such moral code in machines i.e. to articulate the wished values and principles in a way that makes them encoded into machines. This begs the following question: How to incorporate societal values into AVs (or other AI systems)?

Answering this question will require a new approach that can leverage the above-mentioned gap. This approach will involve eliciting the public's expectations, identifying general societal values and moral principles from these expectations about new domains, articulating these values and principles in an operationalizable manner, and, finally, characterizing quantification methods that can help evaluate the performance of these systems, communicate it in an understandable way, and examine its behavior against the expected principles and values. This process will have to be iterative, and it can be painfully slow, but it will be also helpful in other aspects like anticipating public reactions and understand cultural differences.

In this thesis, I describe my experience co-developing a public engagement tool called the *Moral Machine*, which asks people to make decisions about how an AV should behave in various situations. This survey promotes public discussion about the societal and moral values to be embodied by AVs. The survey also allowed us to collect some 30 million decisions that elicit the public's current preferences over these values.

Further, the transition of machines into fully autonomous agents is happening in stages. Certainly, by the time AI systems will take their roles and start making decisions fully independent from any human supervision, these systems would have been experimented with through platforms in which they function in collaboration with humans. Understanding human perception of the agency of machines and judgments about how they share responsibility with humans would provide clear indicators of their reception from the public, once they become in full control. In this thesis, I investigate this question through a study that I co-led.

1.1 RESEARCH QUESTION

This thesis aims to contribute to answering the above-mentioned question: "How to incorporate societal values into AVs (or other AI systems)?". In so doing, I first study how humans attribute responsibility to machines under different levels of automation. Thus, the first question is:

> **Question 1: How do people attribute blame and responsibility to machines under different automation regimes? How does this attribution compare to their judgment of the humans involved in those regimes?**

Performing such type of studies is useful to have an idea about how our attitudes will shape the role of these machines in the future. However, performing studies on such scale might not be enough to achieve the global reach we might hope in order to answer the bigger question above. Thus, a different approach is needed. The second question in this thesis will investigate the possibility for an alternative tool that can substantially scale-up the study to include various factors; present the problem in a simple, engaging, and easy-to-understand way; and to promote public discussion about what societal values we want for machines. Thus, the second questions is:

> **Question 2: How can we elicit the public's judgment over what societal values to embed in machines? How do we promote discussion on such a question among the public?**

1.2 THESIS OVERVIEW

Chapter 2: I provide a background for various terms and topics that relate to the philosophy of ethics, including normative ethics, descriptive ethics and machine ethics. This chapter also calls for the expansion of descriptive machine ethics, a sub-field that is still at its early stage.

Chapter 3: I present a study that investigates how we attribute responsibility and blame to humans and machines in shared-control systems, where both humans and machines work side in side.

Chapter 4: I describe *Moral Machine*, a website that I co-developed for the purpose of collecting humans' judgment over moral decisions made by machines. This chapter also includes some preliminary results that shed some light on potential sources of disagreement over societal principles.

Chapter 5: I conclude this thesis with lessons, limitations, and potential future work that builds on the study in Chapter 3 and that uses the data collected through the platform described in Chapter 4.

2

BACKGROUND

> "A robot may not injure a human being."
>
> –*Asimov*

"Thou shalt not kill" is probably the most straight-forward commandment among the ten, and is arguably the easiest to follow. While clearly stated, this simply-put moral imperative could not have possibly been meant to be absolute, since many teachings of Christianity and Judaism came to include cases for lawful killing as prescriptive imperatives, such as warfare and self-defense. However, interpretations of such cases were never void of debate and disagreement. While Augustine taught that killing in self-defense is a sin [6], Thomas Aquinas later reasoned that killing in self-defense is permissible if killing of the assailer was not intended, even if the killing was foreseeable [4]. With that, Aquinas laid down the foundation for a moral principle that later became a central and renowned concept in ethics under the name of the Doctrine of Double Effect (DDE) [54].

Religion and morality have certainly been closely intertwined, and they both have parts that focus on what one *ought to* do. While intuition is often used when it comes to theorizing what one *ought to* do, a dominant method to test intuition has been employed. In this method, one is engaged in a recurrent, deliberative, dual process of reflection and revision of his/her beliefs about what (s)he *ought to* do with regard to some issue. The end point of this process is named *reflective equilibrium* [64], and we reach it when our general principles and our judgment of specific cases are in agreement. As an example, we may start with a general principle that one should never kill. Now, assume that there is someone attacking you, and your life is at risk. Should you be able to defend yourself, even though this could result in you killing the attacker? In light of this scenario, you might decide to revise your general principle to allow for killing as a result of self-defense (e.g. Aquinas reasoning). On the other hand, someone else might decide to stick to the general principle and refuse to revise it given the conflicting special cases (e.g. Augustine reasoning). The name *reflective equilibrium* is used to describe the whole process (not just the end point), and it is accepted as a coherent model of justification in ethics, most specifically in normative ethics (the part that

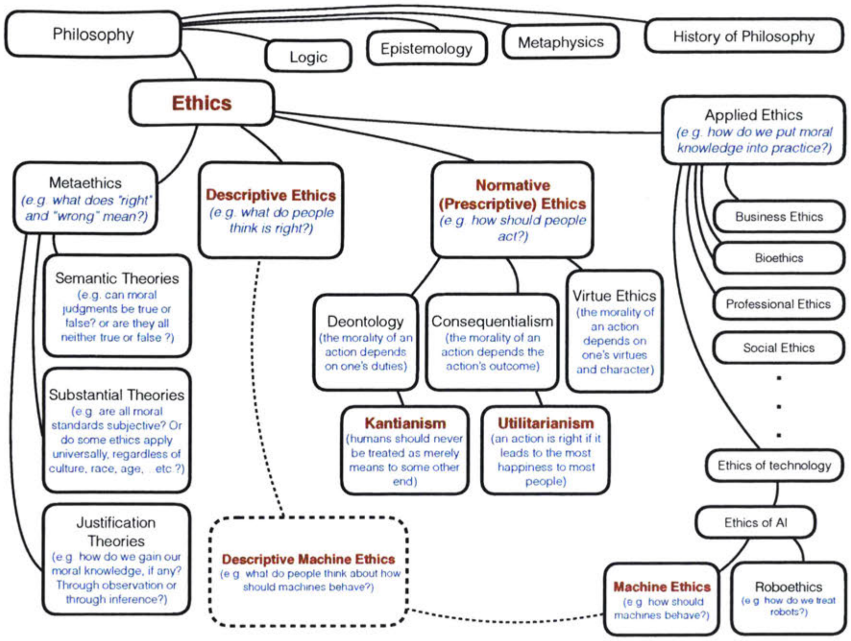

Philosophy
Logic
Epistemology
Metaphysics
History of Philosophy
Ethics
Applied Ethics
(e.g. how do we put moral knowledge into practice?)
Metaethics
(e.g. what does "right" and "wrong" mean?)
Descriptive Ethics
(e.g. what do people think is right?)
Normative (Prescriptive) Ethics
(e.g. how should people act?)
Business Ethics
Bioethics
Professional Ethics
Social Ethics
Semantic Theories
(e.g. can moral judgments be true or false? or are they all neither true or false ?)
Virtue Ethics
(the morality of an action depends on one's virtues and character)
Deontology
(the morality of an action depends on one's duties)
Consequentialism
(the morality of an action depends the action's outcome)
Substantial Theories
(e.g. are all moral standards subjective? Or do some ethics apply universally, regardless of culture, race, age, ..etc ?)
Kantianism
(humans should never be treated as merely means to some other end)
Utilitarianism
(an action is right if it leads to the most happiness to most people)
Ethics of technology
Ethics of AI
Justification Theories
(e.g. how do we gain our moral knowledge, if any? Through observation or through inference?)
Descriptive Machine Ethics
(e.g. what do people think about how should machines behave?)
Machine Ethics
(e.g. how should machines behave?)
Roboethics
(e.g. how do we treat robots?)

Figure 2: An outline of ethics as a subfield of philosophy. This chapter provides a background for parts in red. This figure is not meant to provide a comprehensive outline, but an illustrative one to show how different parts relate. That said, *Descriptive Ethics* is considered by some as not part of philosophy of ethics. Further, the dashed box and links (for *Descriptive Machine Ethics*) are not part of the outline, but added here to illustrate how the main topic of this thesis relates to the rest.

focuses on what one *ought to* do), a subfield of ethics. This explains why thought experiments, ethical paradoxes, and moral dilemmas are crucial components of the study of (normative) ethics, as they provide important tools for the *reflective equilibrium* process.

This chapter provides background about some parts of the field of ethics as a branch of philosophy (see Figure 2 for an outline of the field of ethics).

2.1 NORMATIVE ETHICS

Every action involves three main elements: the action itself, the agent performing the action, and the outcome of this action. When reasoning about what action one *ought to* take in a situation, these three elements become relevant. The main approaches in normative ethics focus on one or more of these three elements.

Also known as prescriptive ethics, the field of *normative ethics* comprises various theories that specify on what basis the morality of an action should be judged. Two main approaches are *deontology* and *consequentialism*. While the latter focuses on the outcome, the former

(a) Jeremy Bentham (1748-1832) [61]

(b) Immanuel Kant (1724-1804) [81]

Figure 3: Jeremy Bentham and Immanuel Kant, leading figures of *consequentialism* and *deontology*, respectively.

focuses on the action, but is also concerned with how it relates to the internal state of the agent. A third main approach, *virtue ethics*, focuses on the agent and their inherent virtues and moral character, arguing that it should be the defining factor of what an agent *ought to* do. The three approaches can result in similar judgments for different reasons. For example, the action to help someone in need would be considered a good action by all three approaches. A consequentialist would consider it good because it brings a good outcome to the one in need, a deontologist would consider it good because the action of helping others is consistent with one's duties towards society, and a virtue ethicist would consider it good because a virtuous person with character traits such as kindness would be inclined to do it. These three approaches are not the only existing approaches; there are other hybrid approaches that combine various elements of each of these three.

2.1.1 *Kantian Deontology*

Consider the following scenario: a criminal who has kidnapped a boy and hid him somewhere is captured. It is crucial to find the boy as soon as possible, as he might be locked somewhere without access to water or food. A thorough search involving helicopters and tracker dogs could not locate the boy's location, and seven hours of interrogation could not bring the criminal to reveal the boy's location.

Imagine that you are the police officer in charge. Your time is limited. You know that if you threaten to torture the criminal, (s)he might crack quickly and give away the boy's location. However, you also know that you have to be prepared to fulfill this threat by hiring "a specialist," and you believe that an individual should not be subjected to torture. Would you threaten the criminal with torture (and be prepared to fulfill this threat)?

If you answered "no" to this question, and you were motivated by the ideal that torture should never be inflicted on any human being, then you are more likely to subscribe to deontological theories than to consequentialistic theories. Deontological theories base morality on our obligation or *duty*. The term *Kantian Deontology* refers to the set of theories, developed by the German philosopher Immanuel Kant, that mainly revolve around *categorical imperatives* – an absolute moral requirement from which all other duties ensue [44]. One formulation of the categorical imperatives is that humans should never be treated as merely means to some end, but always as an end.

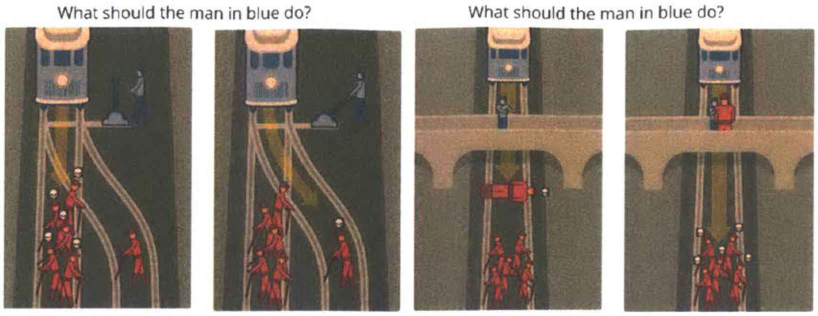
2.1.2 *Utilitarianism as the Consequentialism Paradigm*

Utilitarianism argues that an action is morally permissible if it results in maximizing the happiness and minimizing the pain of the majority. Jeremy Bentham, considered the father of utilitarianism, introduced methods to calculate pleasure and pain [8]. Utilitarianism is the main paradigm of consequentialism. It focuses on the consequences of an action, and, in the broader sense, reduces morally relevant factors to outcomes, which it reduces, in turn, to a single value – *utility*. Even though this can make it seem as simple and shallow, classic utilitarianism underlies various complex claims about the moral rightness of an act.

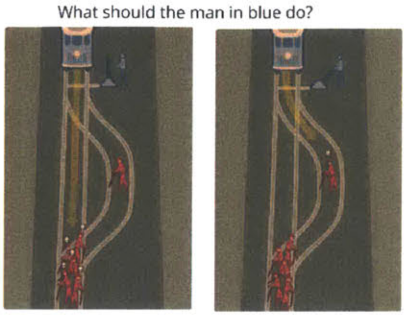
2.2 TROLLEY PROBLEM

With the rise of utilitarianism and deontology, virtue ethics was about to lose its position as the third main approach, if it were not for a few contemporary philosophers who revived the virtue theory. Among these are Elizabeth Anscombe and Philippa Foot, who happened to be close friends with opposing views about contraception and abortion. Both wrote philosophical articles about these matters. One of these articles, by Foot, introduced a version of the Trolley Problem [23]. Later, Judith Jarvis Thomson published two articles with different variants of the Trolley Problem, including the famous version that we know today [76, 77]. The most famous version of the modern Trolley Problem goes as the following (see Figure 4 (a)):

> *"You are standing by the railroad tracks when you notice an empty boxcar rolling out of control. It is moving so fast that anyone it hits will die. Ahead on the main track are five people. There is one person standing on a side track that doesn't rejoin the main track. If you do nothing, the boxcar will hit the five*


What should the man in blue do?
What should the man in blue do?

(a) Spur (b) Footbridge


What should the man in blue do?

(c) Loop

Figure 4: A visual depiction of the *Trolley* Problem – *Moral Machine* interface. (a) The original and most famous version – Spur. (b) A famous variant of the Trolley Problem – Footbridge. (c) Another famous variant of the Trolley Problem – Loop.

> *people on the main track, but not the one person on the side track. If you pull a lever that is next to you, it will divert the boxcar to the side track where it will hit the one person, and not hit the five people on the main track. Which of these two choices is most ethical?"*

As mentioned earlier, the Trolley Problem and its variants provide useful tools for the *reflective equilibrium* process. For the specific scenario above, a utilitarian would probably decide to pull the lever because it would result in saving five people. Another variant of the Trolley Problem features a fat man that can be pushed in front of the Trolley to stop it from killing the other five, but the fat man will die as a result (see Figure 4 (b)). A deontologist would reason in this case that pushing the fat man implies using him as a means to save the five people, and so would prefer to let the five people get killed – something a utilitarian might not approve of.

2.3 DESCRIPTIVE ETHICS

The scenario mentioned in Section 2.1.1 about the kidnapped boy is in fact a true story [86]. The kidnapping happened in Germany in 2002. The kidnapper was a law student in their mid-twenties, and the boy was the son of a banker. What did the police officer in charge do? After seven hours of interrogation and with the failure to locate the boy, the officer instructed interrogators to threaten the criminal. Meanwhile, a specialist was put on call, and was prepared to inflict unimaginable pain on the criminal without any lasting physical damage. The threat was very influential; the criminal confessed on the location, but it was found out that he had already killed the boy. As a result, Frankfurt's deputy police chief was charged with coercion.

This story sparked a public debate on police employing torture. While the officer found support among the German public, being called a "tragic hero" by some [46, 88], he was also criticized by others [9]. There were expressions of outrage from both sides. A public survey in Germany showed that 60% of participants thought that the officer should not have been charged [88]. If we are to accept this survey, it shows that the majority can disagree with what law professors, philosophers, and experts generally think is sensible. This is definitely not the first incident to illustrate this, but it provides another motivation for the field of *experimental philosophy* [2] that *descriptive ethics* is part of.

While normative ethics focuses on how people should act, descriptive ethics focuses on what people think about how they should act. Articles in philosophy usually contain unqualified statements about what a person with "sound" views would think in a specific scenario. Anecdotes and examples are sometimes employed to support the various claims. However, descriptive ethics, while still counted by many philosophers as "adding nothing to philosophy," has uncovered many interesting findings, such as cross-cultural intuitions [37, 84], effect of demographic variability [24, 41] and cognitive biases [47, 57, 80]. Further, it helps diagnose society's biases, and predict potential public outrage before a given law is enacted.

The Trolley Problem has been often used in descriptive ethics to extract the public's moral decisions. Interestingly, according to many studies, while most people would pull the lever to save five people, most are not willing to push the fat man in front of the trolley to save five people [21, 32]. Surveys including Trolley Problem have been also used in moral psychology, cognitive science, and neuroscience [30, 31, 80]. Probably the most significant empirical result involving the Trolley Problem was found by Green et al. [30, 32]. They found, using fMRI, that some decisions that are rather personal (such as pushing the fat man) are associated with engaging emotional regions in the brain. In contrast, they found that other less personal decisions (such as pulling the lever) are associated with engaging regions of the brain associated with reasoning.

2.4 MACHINE ETHICS

Discussions about machine ethics have been feeding on science fiction to some degree. Modern articles on machine ethics still include a reference to Asimov's laws (AKA three laws of robotics), which were introduced in one (and some other) of his stories "Runaround" [5]:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

A Zeroth law was included later by Asimov to prioritize saving humanity over saving a human being. Other science fiction authors also had their share in adding some other laws including ones that require robots knowing that they are robots, and robots reproducing [34], among other laws.

Such laws may be useful as a first step in establishing some moral principles and concepts for autonomous machines. However, their ambiguity, inapplicability, and misrepresentation have been questioned [3]. Further, it is unclear how such laws will be embedded in robots. The same issue arises when trying to embed Kant's categorical imperative [62]. Bostrom and Yudkowsky have argued for the use of *decision tree algorithms* (e.g. *ID3*) to encode principles [12]. However, the common trend seems to be heading towards machines learning their behavior from observation and examples as occurs in *artificial neural nets*. While using the decision trees approach has the advantage (in comparison to the learning approach) of producing machines with transparent, clear, and predictable behavior, the fact that human ethics are rather complex, evolving, and hard to operationalize in their current form make a better case for the learning approach.

Outside of science fiction, there have been different arguments that either defend or criticize the possibility of machines becoming moral agents. Arguments were mainly brought about whether machines would ever have intentions and freedom, and whether this will qualify them as ethical agents [22, 39, 43]. Others argued that the mere fact that machines can produce outcomes with ethical implications would make them "ethical entities." The "ethical impact" of machines was the main focus of Moor when he defined four types of ethical machines (listed in increasing levels of ethical agency) [55]:

1. Ethical impact agents: Machines that have (un)intended ethical consequences (e.g. your fitness app that encourages you to work out).

2. Implicit ethical agents: Machines that employ some features to retain safety, security, or convenience (e.g. email spam filtering algorithms).

3. Explicit ethical agents: Machines that have general principles embedded that it uses to make ethical judgments. These could be a realistic goal for ethical machines in the near future.

4. Full ethical agents: Machines that are ethical in the same level that humans are ethical. These machines have free will, consciousness, and intentions.

These increasing levels of ethical agents correspond to increasing levels of autonomy, user risk, and directness of the relationship between the user and the machine. In turn, these increasing levels of moral agency were argued to correlate with increasing levels of user trust in these agents [74].

Wallash and Allen [83] proposed an alternative classification of three moral stages in AI, based on *autonomy* and *sensitivity to values*: 1) operational morality, which describes machines lacking autonomy and ethical sensitivity but still employ some moral values (this corresponds to the *implicit ethical agents* level); 2) functional morality, which describes machines that have high autonomy and low ethical sensitivity, or low autonomy and high ethical sensitivity (for example, F-16 autopilot and decision support system for doctors, respectively); and 3) full moral machine, which describes machines with high autonomy and high sensitivity (this corresponds to *full ethical agents* level).

Until recently, talking about machines that have "explicit ethical agency" seemed a rather futuristic topic. However, the advent of autonomous vehicles (AVs) has prompted multidisciplinary discussions about the moral implications of decisions taken by these AVs. Most of these discussions revolved around moral dilemma situations. Lin [48, 50] argued that the discussion of ethics will become a necessity for AVs. Similarly, Goodall [27, 29] argued that AVs will sooner or later have to make decisions with ethical consequences. Both Lin and Goodall proposed various Trolley-like no-win scenarios that the car might face, and addressed the potential criticism of the possibility of these scenarios to happen. Later, Goodall [28] reframed the issue as a risk management problem in which the car faced scenarios with crash risk and uncertainty.

2.5 TOWARDS DESCRIPTIVE MACHINE ETHICS

Given that the prospect of self-driving vehicles being delegated the resolution of moral dilemmas still seems like an unrealistic idea, only a few studies have explored the public's opinion regarding what kind of ethical decisions we want these autonomous machines to make. However, these very few studies have already provided some insightful outcomes. For example, when surveying people about the possibility of an autonomous vehicle sacrificing a passenger to save one or more pedestrian lives, Bonnefon et al. [11] discovered that there are social implications for this type of question. While respondents thought it a good idea for an autonomous vehicle to be utilitarian between passenger and pedestrian lives in principle, they themselves would not want to own such a vehicle. This highlights a potential social dilemma and a new *tragedy of the commons* in the context of driverless cars [33]. In another study, Malle et al. [53] found that when making a non-utilitarian decision, a robot driver will be blamed more than a human driver. Such studies are paving the way for a new subfield that we can call "Descriptive Machine Ethics."

3

MORAL RESPONSIBILITY ACROSS LEVELS OF VEHICLE AUTOMATION

> Full autonomy is a wonderful goal. But none of us in the automobile or IT industries are close to achieving true Level 5 autonomy.
>
> –*Gill Pratt*

Autonomous machines such as self-driving vehicles may ultimately make decisions with moral implications, such as deciding who to kill in the event of an unavoidable accident [11]. Psychological studies suggest that these autonomous machines will be perceived as moral agents, who deserve a measure of responsibility and blame for the consequences of their actions [51–53, 68]. Autonomous machines which are perceived as blameworthy for unethical outcomes may suffer significant setbacks with respect to their adoption and social acceptance. This problem is compounded by the fact that in many domains, full autonomy will only be gradually attained, by going through phases of shared control between humans and (increasingly autonomous) machines. Accordingly, to ensure a smooth transition to full autonomy, we must understand how public opinion will allocate responsibility and blame between human and machine, when things go badly in a situation of shared control. Here we map the allocation of responsibility and blame between human drivers, autonomous vehicles (AVs), their programmers, and their manufacturers, for various levels of automation, after accidents in which five pedestrians were killed when another trajectory would have killed only one. We show that the machine is blamed the most, both in relative and absolute terms, under the early stages of automation that we are currently experiencing. Accordingly, we argue that the most critical period for the social acceptance of automated driving is *right now,* and that we can no longer defer the discussion on its ethical implications.

Consider an accident in which a vehicle crashed and killed five pedestrians, while another trajectory would have killed only one pedestrian. If the vehicle was fully autonomous, its decision to kill five pedestrians when it could have killed a single one would be perceived as unethical by public opinion [11]. The user of the vehicle, not being involved in this decision, would likely be exonerated from responsibility and blame. Instead, the blame would then fall on the 'industry': the company that produced the autonomous vehicle (AV), the programmers who designed its code, or even the AV itself. Given the millions of miles that will be driven by tens of thousands of AVs, even the small chance of AV programming making an unethical decision will no doubt lead to well-publicized cases. Such publicity would in turn degrade the image of AVs, slow down their adoption, and jeopardize their promised safety and environmental benefits.

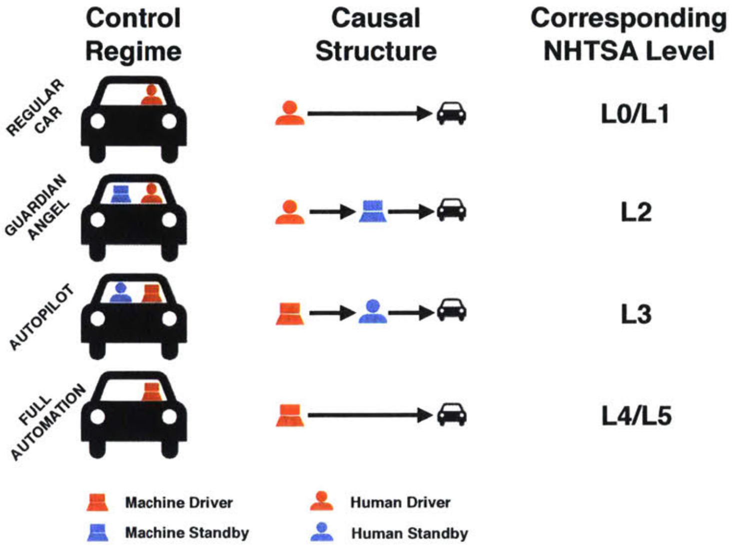


Figure 5: Four levels of automation for autonomous vehicles: Regular car (RC), Guardian Angel (GA), Autopilot (AP), and Fully Autonomous (FA) regimes. These levels broadly correspond to the classifications of the NHTSA and SAE International [79]. Levels 1 (RC) and 4 (FA) involve only one driver and no standby. Levels 2 (GA) and 3 (AP) involve two agents: the main driver (shown in red), and the standby driver (shown in blue). The causal structure of the decision making is different across the four levels, but in each case, the main driver makes a decision (in this article, stay or swerve), and the standby driver (if any) then decides whether to override the decision of the main driver.

Fully autonomous AVs, though, are still far in the future. While the technology needed for their deployment is advancing rapidly, what we are currently witnessing is a gradual increase toward full automation, going through several steps of shared control between user and vehicle (Figure 5). Some vehicles can take control over the actions of a human driver (e.g., Toyota's 'Guardian Angel') to perform emergency maneuvers. Other vehicles may do most of the driving, while requiring the user to constantly monitor the situation and be ready to take control (e.g., Tesla's 'Autopilot'). Before we reach full automation, public opinion about AVs will be shaped by the way people allocate responsibility and blame between user and industry, following suboptimal outcomes that were the result of shared control.

In this article, we consider a simplified model of shared control in which a suboptimal outcome (five pedestrians died where only one could have died) resulted either from (a) a decision to override the maneuver initiated by the main driver; or (b) a failure to override the maneuver initiated by the main driver. We consider in turn two questions. First, how do automation regimes affect the *relative* allocation of responsibility and blame between human and machine? Second, how do automation regimes affect the *absolute* allocation of responsibility and blame to human and machine?

There is no obvious response to the first question, since people can shift responsibility and blame either toward the agent who contributed the most to the outcome, or to the agent who had the last opportunity to act [15, 18, 26, 72, 87]. In case of a suboptimal outcome under the Guardian Angel Regime, the user does most of the driving, but the decision to override (and thus to act last) pertains to the machine. In contrast, under the Autopilot regime, the machine does most of the driving, but the decision to override pertains to the user. If people are especially sensitive to this latter fact, they may (for example) blame the human user for not overriding a suboptimal course, more than they blame the Autopilot for setting up that course in the first place—leading to the paradoxical result that users are blamed *more* when they are *less* in control. Such a result would have important implications for the adoption of AVs, since it would suggest that increasing automation might progressively shift the blame away from AVs, and thus progressively reduce the public impact of suboptimal crash outcomes.

The question about *absolute* levels of blame derives from the observation that responsibility and blame are not necessarily conserved when several agents contribute to a single outcome [60]. In other words, we cannot assume that, in situations of shared control, the blame assigned to the human user will decrease in direct proportion to the blame assigned to the machine. This creates the possibility that the absolute amount of blame assigned to the human user may be greater when control is shared than when the human makes all the decisions—and conversely, that the absolute amount of blame assigned to the machine may be greater when control is shared than when the machine makes all the decisions. Accordingly, we cannot just identify the automation regimes that shift the balance of blame from human to machine, or vice versa—we also need to identify the automation regimes that lead to the greatest absolute judgments of blame to humans and machines. Only with this full picture will we be able to anticipate the ups and downs of public opinion about AVs during the transition period toward full automation.

3.1 RESULTS

We presented 973 participants with multiple crash scenarios that all ended with a suboptimal outcome: five pedestrians were killed, where another trajectory would have killed a single one. In half the scenarios, this outcome resulted either from a suboptimal action (the vehicle was headed toward one pedestrian, but swerved into five); in the other half, the outcome resulted from a suboptimal inaction (the vehicle was headed toward five pedestrians, and stayed on course instead of swerving into one). Furthermore, different scenarios featured the different levels of automation shown in Figure 5.

In other words, in scenarios featuring Regular Cars and Full Automation, the driver (respectively man or machine) crashed in the five pedestrians after a suboptimal action or inaction. In shared control scenarios, either the main driver headed toward one pedestrian and the standby driver took the suboptimal action of swerving into five, or the main driver headed toward five pedestrian and the standby driver took the suboptimal action of staying on course.

For each scenario, participants rated the causal responsibility and blameworthiness of the human user (each on a scale from 0 to 100), as well as the causal responsibility and blameworthiness of the industry (each on a scale from 0 to 100). The term 'industry' itself never appeared in the questions, though. For some participant the questions were about the company that produced the car; for others, about the person who programmed the car; and for yet others, about the car itself. Below, we will report both aggregated results (in which these three agents are collapsed under the catchall term 'the industry') and detailed results (distinguishing between these three agents).

3.1.1 *Relative allocation of responsibility and blame*

Figure 6 summarizes the *relative* allocation of responsibility and blame between user and industry, after a suboptimal action or inaction, for the four automation regimes, by displaying the 95% confidence interval of the raw difference between ratings given to the industry and ratings given to the user. Confidence intervals that only contain positive values indicate greater responsibility and blame for the industry, and confidence intervals that only contain negative values indicate greater responsibility and blame for the user.

Unsurprisingly, the industry is allocated more responsibility and blame when the crash featured a fully autonomous car. Just as unsurprisingly, the user is allocated more blame than the industry when the crash featured a regular car. More interesting for our current purpose are situations of shared control.

Strikingly, the user is blamed more than the industry (and perceived as more responsible than the industry) under the Autopilot

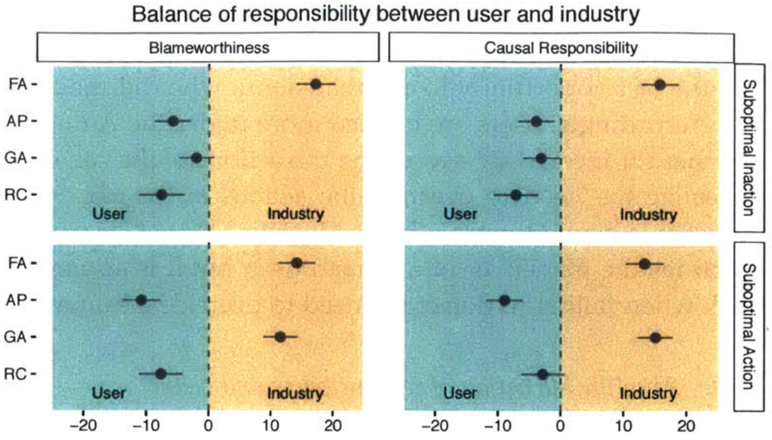


Figure 6: 95% confidence intervals of the difference in ratings of causal responsibility and blameworthiness between the Industry and the User, across four automation regimes, after a suboptimal action or inaction.

regime, in which the car actually does most of the driving. This is true for suboptimal actions (where the user overrides the car and swerves into five pedestrians), but also for suboptimal inactions (where the user fails to override the car which is headed toward five pedestrians). It is also true regardless of the specific agent standing for the industry (car, programmer, company). Conversely, the industry is blamed more than the user (and perceived as more responsible than the user) under the Guardian Angel regime, but only when the car incorrectly overrides the decision of the user (regardless of the specific agent standing for the industry). When the car fails to override the action of the user, respondents assign broadly the same ratings of responsibility and blame to the user and the industry (albeit directionally greater ratings to the user).

These results are confirmed by a mixed-model analysis in which the dependent variable was the difference between the rating given to the industry and the rating given to the user; the fixed effects were the agent standing for the industry (car, company, or programmer), the nature of the suboptimal decision (action or inaction), and the automation regime (RC, GA, AP, or FA); and participants were entered as a random factor. When conducted on the blame responses, this analysis detected a significantly negative intercept term ($p < .005$), reflecting more overall blame for the user. The analysis detected two exceptions to this general trend: the industry was blamed more in the FA regime ($p < .001$), and in the GA regime ($p < .001$), although this last exception only held for suboptimal actions, as reflected in a significant interaction term ($p < .001$). Results were exactly similar when the analysis was conducted on causal responsibility ratings (full details are provided in the Appendix).

To summarize our findings on the relative allocation of responsibility and blame, it appears that respondents care more about who had the last opportunity to act, than about who did most of the driving. Accordingly, users are blamed more under the Autopilot regime, whether for incorrectly overriding the actions of the car, or for failing to realize the need to override the actions of the car. Similarly, the industry is blamed more when a Guardian Angel system incorrectly overrides the maneuver of a human user; but it is apparently off the hook when failing to detect the need to override a human maneuver.

3.1.2 *Absolute allocation of responsibility and blame*

Figure 15 displays the *absolute* allocation of responsibility and blame to the user and the (agents standing for the) industry, for the four levels of automation. As shown in Figure 15, the user received the highest ratings of responsibility and blame under the Autopilot regime (in which control is shared), rather than in the Regular Car regime (in which the user makes all the calls). Similarly, the company that produced the car received the highest ratings of responsibility and blame under the Guardian Angel regime (in which control is shared), rather than in the Fully Autonomous regime (in which the car makes all the calls).

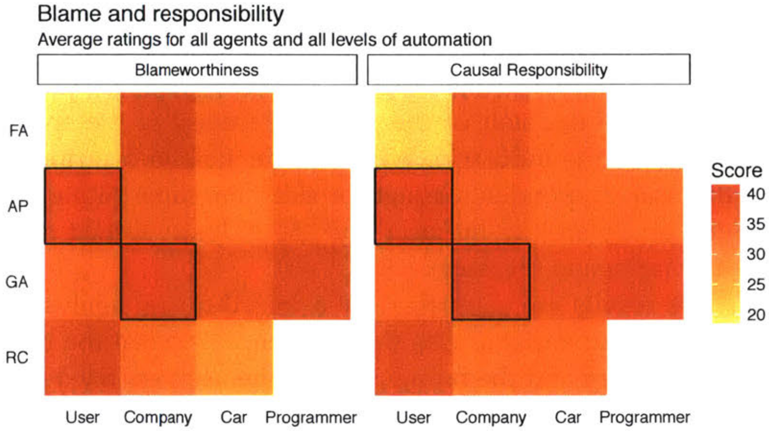


Figure 7: Causal responsibility and blame-worthiness ratings of user, car, programmer, and company. Values for programmer in regular cars (RC) and fully-autonomous (FA) were not collected. The user receives highest ratings under the Autopilot (AP) regime, while the company receives highest ratings under the Guardian Angel (GA) regime.

These results were broadly confirmed by mixed-model analyses in which ratings of responsibility and blame were entered as the dependent variable, automation regime was entered as a fixed effect, and participants were entered as a random factor. The analyses of users' ratings used the Autopilot regime as the reference level of automation. Blame ratings at this reference level were higher than at any other level ($p < .001$ against FA, $p < .001$ against GA, $p = .01$ against RC). Responsibility ratings at this reference level were higher than at the FA level ($p < .001$) and the GA level ($p < .001$), but only directionally higher than at the RC level ($p = .31$). The analyses of company's ratings used the Guardian Angel regime as the reference level of automation. Responsibility ratings at this reference level were higher than at any other level ($p < .001$ against RC, $p < .001$ against AP, $p = .014$ against FA). Blame ratings at this reference level were higher than at the RC level ($p < .001$) and the AP level ($p < .001$), but only directionally higher than at the FA level ($p = .22$).

In sum, it appears that blame and responsibility are not conserved in situations of shared control between human and machine. Users are blamed the most under the Autopilot regime, more than under the Regular Car regime; and companies are seen as most responsible under the Guardian Angel regime, more than under the Full Automation regime. On the basis of these results on the absolute and relative allocation of responsibility and blame across automation regimes, we are in a position to sketch the likely evolution of public opinion about self-driving cars during the transition period toward full automation.

3.2 DISCUSSION

In a ideal world, fallible human drivers would gradually cede control to their car, whose automated driving would make everyone safer. Human drivers would build trust in automated driving thanks to their introduction to Guardian Angel systems, and this trust would prepare them to take the backseat and leave the driving to Autopilot systems. Once used to watch their car drive itself in Autopilot under their monitoring, humans would be ready to cede total control and accept Full Automation.

What makes this scenario unrealistic is that it neglects the impact of the crashes that will take place in situations of shared control. Some accidents will not be avoided, some lives will be taken, and someone will have to take the blame for them. Any loss of lives that is blamed on a self-driving car (or the company that produced the car) may create public backlash, endanger the social acceptance of automated vehicles, and slow down their adoption.

According to our results, the greatest point of fragility in the 'gradual automation' scenario is right now, as many drivers are being introduced to Guardian Angel systems – a fragility that may become especially apparent if and when accidents based on such systems (or even just accused of being based on such systems) begin to emerge. Indeed, we found that the industry would be blamed the most for lives taken by the intervention of a Guardian Angel system. In fact, participants judged that the responsibility of the industry was even greater in this case than for the fully autonomous vehicles that will eventually make all the driving decisions, without human supervision. Importantly, we found that the Autopilot stage was not as dangerous (in term of public backlash) than the Guardian Angel stage that precedes it. Even though cars at the Autopilot stage do most of the driving, the human supervisor takes the blunt of the blame when lives are unduly taken. In fact, the human supervisor of a car on Autopilot receives even more blame than the human driver of a regular car.

These results are both informative and provocative. Informative, because they tell us that now is the time we need to address the ethical issues raised by automated driving. According to our findings, the ethical implications of Guardian Angel decisions could make or break public trust in automated driving. Therefore, we cannot defer ethical discussions until after fully automated vehicles are ready to roll, because the public acceptance of these vehicles is being shaped now, through the actions of Guardian Angel systems.

Our results are also provocative, because they point at a paradox in the way people think of sharing control with machines. Consider for a moment that people blame the human supervisor, more than the Autopilot machine, after a suboptimal death toll. That is, they blame the human supervisor for a bad decision that took place in a split-second, more than they blame the very expensive, very sophisticated, very fast machine whose reason to exist is to make better decisions than humans. It would appear that no matter how sophisticated the machines that we build to decide for us, we still blame ourselves for the bad decisions they make under our watch—and that we blame ourselves even more than if we actually made the bad decision.

It is urgent to understand this phenomenon, since it has significant implications on both the regulatory approaches that assign blame and liability, and on insurance products designed to hedge risk against such liability. Indeed, our biases in assigning blame may lead to suboptimal insurance models that do not capture *actual* statistical risk, but are skewed by *perceived* risk.

Indeed, the belief that 'humans should know better' is a double-edged sword: as much as it can make us forgive the mistakes made by machines, it can also make us underestimate their capabilities. It might be that for people to take full advantage of intelligent machines, they first need to make the machines accountable for their mistakes.

3.3 METHODS

The data was collected between September and November 2016 from 973 participants. The study was programmed on Qualtrics survey software and participants (USA residents only) were recruited from the Mechanical Turk platform, and each was compensated with 35 cents. Participants were presented with crash scenarios that resulted in a suboptimal outcome. Each scenario featured one of the following types of car: 1) single-control cars: which have one driver, and 2) shared-control cars: which have a main driver (who is in control at the start of the scenario) and a secondary driver (that can intervene). After each scenario, participants were asked to attribute causal responsibility and blameworthiness to two agents: the human in the car (the user); and a representative of the car, being it the car itself, the programmer of the car, or the manufacturing company of the car. The representative of the car (car vs. company, for single-control; and car vs. programmer vs. company, for shared-control), and whether the control is shared (single-control vs. shared-control) were manipulated between subjects.

Each participant read four vignettes in which a car is being driven by either a single driver (single control) or two drivers (shared control). Drivers are faced with a tradeoff between killing five pedestrians and killing one pedestrian. The final outcome of the decisions made by drivers is the death of the five pedestrians. In the case of shared control, the main driver always makes a decision as to keep the car on its track (no intervention). The four vignettes per participant manipulated the identity of the drivers (i.e. in the single-control case whether the driver is a human or a machine, and in the shared-control case whether the main driver is a human, and the secondary is the machine, or vice versa), and the decision by the last driver (that is whether the single-control driver swerved or stayed; and whether the shared-control secondary driver overrode or did not override). Next, participants were asked to indicate whether each of the human and a representative of the car is "not blame-worthy" or "very blameworthy", and to indicate to what degree each of these two agents has caused the death of the five people (from "very little" to "very much") all on 0-100 sliders anchored at these two expressions for each question. This makes up to four questions per vignette. The order of vignettes and questions were all counterbalanced. When the car was driven by the human only, it was referred to as "car". In all other cases, the car was instead referred to as "robocar", and was described as a state-of-the-art self-driving car. At the end of the survey, participants provided basic demographic information (e.g., age, sex, income, political views).

To get our final sample (of 973 participants), we started with all responses to our online surveys (which can be greater than the number of responses requested on Amazon Turk if some subjects complete the survey, but do not indicate on Amazon Turk that they have completed it). We then excluded any subjects who did not (i) complete all measures within the survey, (ii) transcribe (near-perfectly) a 169-character paragraph from an image (used to exclude non-serious

Turkers), and (iii) have unique TurkID per study (all records with a recurring MTurk ID were excluded).

4

MORAL MACHINE AS A MASSIVE ONLINE EXPERIMENTATION (MOE) TOOL

> "The Trolley Problem may be overused,
> but this 'moral machine' from MIT is fascinating."
>
> –*Erik Brynjolfsson*

As mentioned earlier in Chapter 2, there had been no extensive empirical studies on human moral perception of moral decisions made by autonomous vehicles, and only a few of the possible factors were considered.

Indeed, a large number of factors apart from mere utilitarianism would have to be considered when determining the appropriate outcome of a moral dilemma. One is the relationship of the persons involved to the vehicle: are they passengers of the vehicle, or are they pedestrians and passers-by?

Another factor to consider is whether a given outcome involves taking action (commission) or not taking action (omission). Shallow et al [70], for example, find that omission was strongly preferred to commission when the decision between moral dilemma outcomes was made by humans.

Yet another variable could be the traffic laws. Is a jaywalking pedestrian or a pedestrian crossing at a pedestrian "wait" signal the moral equivalent of a law-abiding pedestrian crossing at a "walk" signal? Should a passenger be sacrificed if one or more pedestrians are ignoring crossing signals, but not if they are abiding by them? Would it be acceptable to seriously injure a law-abiding pedestrian to save a law-flouting pedestrian's life?

And finally, do age, gender, fitness level, and social status factor into decisions to save or kill, to the extent that these can be determined, and to the extent they might factor into perceived survival likelihood and/or value to society?

In order to account for the possible combinations of factors and dimensions, we required a system that could generate a large number of scenarios under constraints that would keep the scenarios realistic, and have large numbers of users look through and vote on acceptable outcomes in multiple scenarios. This would be extremely difficult to achieve using traditional crowd-sourced experimentation platforms, and – even scaled down – prohibitively expensive.

Fortunately, a new type of experiments has emerged that can permit conducting studies with large-scale of participants, over exponential number of conditions, within short time and free of charge. Further, unlike observational big data, this type of experiments allow for control on the experimenter side, and provide the possibility for random allocation. This new type is called Massive Online Experiments (MOEs).

4.1 MASSIVE ONLINE EXPERIMENTS (MOE)

Massive online experiments (MOE) are a special type of *web-based experiments*, experiments conducted over the Internet. MOEs usually target massive sample sizes (e.g. hundreds of thousands or millions of users), and are usually either conducted through online social networks such as Facebook [10], or through web-sites, or services that are designed specifically to attract diverse public users [19].

Unlike lab-based experiments, field experiments, and natural experiments, *web-based experiments*, have the advantages of recruiting larger sample pools, of more diverse background, within a shorter period, and at a lower or no cost [66]. MOEs enjoy all of these advantages to a higher degree, in addition to the possibility of conducting cross-cultural studies effectively.

On the other hand, web-based experiments (and MOEs) are criticized for suffering from some shortcomings. One of the main shortcomings is the difficulty to control the conditions in which users are taking the surveys, which makes it hard to prevent manipulation by users e.g. users can take the same test multiple times, users can take the survey less seriously, or users can provide misinformation that are hard to verify. Other limitations include the self-selectivity of users in joining and dropping out of the experiments. Further, users might have concerns about their privacy and anonymity when doing such experiments, which might influence their answers. A criticism for cross-cultural samples is the non-representativeness of English-speaking users coming from non-English speaking countries.

Despite these limitations, experiments conducted through the web has been successfully replicated through other types [38, 65, 69]. Further, one can argue that some of these criticisms can hold for other types of experiments, as well.

4.2 MORAL MACHINE

In order to achieve the goal of collecting, analyzing, and studying the different factors that are relevant to the moral judgment made by machines, we built the Moral Machine, a platform for gathering data on human perception of the moral acceptability of decisions made by autonomous vehicles faced with choosing which humans to harm and which to save. Moral Machine fits the specifications of a massive online experimentation tool given its scalability, accessibility to online community, and the random assignment of users to conditions.

Another purpose to the project was the facilitation of public feedback and discussion of scenarios and acceptable outcomes, and especially public discussion of the moral questions relevant to self-driving vehicles, which was greatly lacking. Following, I describe the requirements we set for the Moral Machine, the implementation process, and the different interfaces that this platform offers, before I end this chapter with preliminary results from the collected data.

4.2.1 *Requirements*

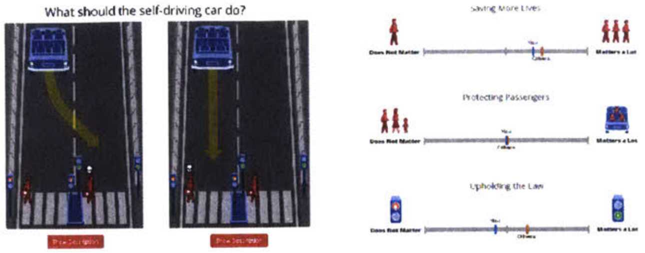
The success of MOEs is contingent on the virality of the platform i.e. its success to attract so many users. This is not an easy goal to achieve, given that most successful online platforms owe their popularity to various different factors, some of which are context-dependent, and others are beyond control or are unclear apriori. However, some common practices are usually suggested. For example, a platform that aims to attract hundreds of thousands of users (or millions of users) need to provide a clean, tight user experience that the user would feel compelled to share with their social network, the experience would have to be easy to enter, be short, and be visually engaging, so as to put as many users as possible through as many combinations of scenarios as possible. It would also need to hold up a "personality mirror" to the user by summarizing their responses, and showing them how they compare to others.

To achieve the second goal of the project (i.e. promoting the public discussion), we required the platform to have features that would permit users to assemble their own scenarios, to view others' scenarios, and to provide feedback and engage in discussion on both.

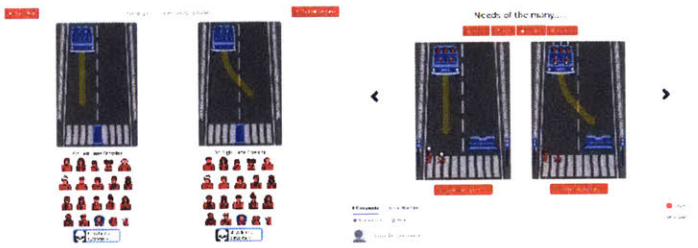
In order to encourage organic propagation of awareness of the platform, and thereby gather more data through it, it would also have to include features that make it easy to share on social networks – specifically, a scenario the user themselves might have assembled, any interesting scenarios a user might come across while browsing the gallery, as well as the user's own performance summary.

4.2.2 *Implementation*

Meteor was chosen as the development platform, for its responsiveness, useful packages, dynamic scripting, and template-based structure. The application was developed using a rapid prototyping methodology, and deployed on a cloud application hosting service with a sep-


What should the self-driving car do?
Saving More Lives
Does Not Matter
Matters a Lot
Protecting Passengers
Does Not Matter
Matters a Lot
Upholding the Law
Does Not Matter
Matters a Lot

(a) Judge mode (b) Section of User Results


Needs of the many...

(c) Design mode (d) Browse mode

Figure 8: *Moral Machine* interface.

arate accelerated web hosting service. It is optimized for social media sharing with Cards, Markup, and Open Graph tags, features prominent calls to action on the main page, and is developed to be responsive for usability on mobile devices. The site's main page features a video describing the project, and offers instructions and background information to view, while linking to three modes of user experience: Judge, Design, and Browse.

4.2.2.1 *Judge*

The central data-gathering feature is the Judge mode, seen in Fig. 8 (a). In this mode, users are presented with a series of 13 moral dilemma scenarios, with a simple point-and-click (or, in the case, of the mobile version, toggle-and-commit) method to choose which outcome of the two possible for a given scenario was deemed by the user to be most acceptable.

Each scenario features characters from the following set: C = {*Man, Woman, Pregnant Woman, Stroller, Elderly Man, Elderly Woman, Boy, Girl, Homeless Person, Large Woman, Large Man, Criminal, Male Executive, Female Executive, Female Athlete, Male Athlete, Female Doctor, Male Doctor, Dog, Cat*}.

The scenarios are generated using randomization under constraints chosen so that each scenario tests specifically for a response along one of the following given dimensions:

1. **Species.** This dimension tests the extent to which users are willing to save/sacrifice pets when put against humans. We con-

sider two sets of characters: 1) pets: $S_1 = \{\mathrm{Dog}, \mathrm{Cat}\}$, and 2) humans: $S_2 = C \backslash S_1$. To generate a scenario of this dimension, the number of characters on each side[1] (same number on both sides) z is sampled from the set of positive integers less than 6 . Then, z pairs of characters are sampled (unordered sampling with replacement) from the Cartesian product of the two sets $S_1 \times S_2$. The first entries of the ordered pairs (i.e. pets) go to one side, while the second entries of the ordered pairs (i.e. humans) go to the other side.

Given this,[2] the number of distinct scenarios of this dimension is $\sum_{i=1}^{5}\left[\binom{x_1+i-1}{i}\binom{x_2+i-1}{i}\right]$, where $x_1 = |S_1| = 2$, and $x_2 = |S_2| = 18$. Hence, the number of distinct scenarios of this dimension is $N_{Species} = 193,038$.

2. **Social Value.**[3] This dimension tests the extent to which users are willing to save/sacrifice characters of higher social value (e.g. a *Pregnant Woman*) when put against characters of lower social value (e.g. a *Criminal*). We consider three sets of characters, corresponding to three levels: 1) characters of low social value: L_1 = {*Homeless Person, Criminal*}, 2) characters of neutral social value: L_2 = {*Man, Woman*}, and 3) characters of high social value: L_3 = {*Pregnant Woman, Male Executive, Female Executive, Female Doctor, Male Doctor*}. To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers less than 6. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set: $(L_1 \times L_2) \cup (L_1 \times L_3) \cup (L_2 \times L_3)$. The first entries of the ordered pairs (i.e. lower-level characters) go to one side, while the second entries of the ordered pairs (i.e. higher-level characters) go to the other side.

   Given this, the number of distinct scenarios of this dimension is

$$\sum_{i=1}^{5}\sum_{j=0}^{i}\left[\binom{x_1+j-1}{j}\binom{x_2+i-j-1}{i-j}\binom{x_2+x_3+j-1}{j}\binom{x_3+i-j-1}{i-j}\right]$$

   where $x_1 = |L_1| = 2$, $x_2 = |L_2| = 2$, and $x_3 = |L_3| = 5$. Hence, the number of distinct scenarios of this dimension is $N_{SocialV} = 58,547$.

---

[1] We use the term *side* to refer to one of the two options that the cars will choose to save/kill. Depending on the *relationship to vehicle* dimension (mentioned later), the *side* can refer to inside the car, or on the zebra crossing ahead or on the other lane.

[2] Note that in all cases we do unordered sampling with replacement. Hence, the formula $\binom{n+k-1}{k}$.

[3] Note here that "social value" refers to the *perceived* social value i.e. the widespread perception of the used characters. We do not endorse the valuation of any humans above others. With that being said, we do not suggest that AVs discriminate on the basis of any of the classifications presented in Moral Machine.

3. **Gender.** This dimension tests the extent to which users are willing to save/sacrifice female characters when put against male characters. We consider two sets of characters: 1) female characters: $G_1 = \{$*Woman, Elderly Woman, Girl, Large Woman, Female Executive, Female Athlete, Female Doctor*$\}$, 2) male characters: $G_2 = \{m \mid m = g(f), f \in G_1\}$, where g is a bijection that maps each female character to its corresponding male character (e.g. $g($*Female Athlete*$) =$ *Male Athlete*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers less than 6. Then, z pairs of characters are sampled (unordered sampling with replacement) from: $\{(f, m) \mid f \in G_1, m = g(f)\}$. The first entries of the ordered pairs (i.e. female characters) go to one side, while the second entries of the ordered pairs (i.e. male characters) go to the other side.

   Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |G_1| + 1 = 8$. Hence, the number of distinct scenarios of this dimension is $N_{Gender} = 791$.

4. **Age.** This dimension tests the extent to which users are willing to save/sacrifice characters of younger age when put against characters of older age. We consider three sets of characters, corresponding to three levels: 1) characters of young age: $A_1 = \{$*Boy, Girl*$\}$, 2) neutral adult characters: $A_2 = \{$*Man, Woman*$\}$, and 3) elderly characters: $A_3 = \{$*Elderly Man, Elderly Woman*$\}$. Consider the following two gender-preserving bijections $a_1 : A_1 \to A_2$, and $a_2 : A_2 \to A_3$ (e.g. $a_1($*Boy*$) =$ *Man*, and $a_2($*Woman*$) =$ *Elderly Woman*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers less than 6. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set:

$$\{(y, n) \mid y \in A_1, n = a_1(y)\} \cup$$
$$\{(n, d) \mid n \in A_2, d = a_2(n)\} \cup$$
$$\{(y, d) \mid y \in A_1, d = a_2 \circ a_1(y)\}$$

   The first entries of the ordered pairs (i.e. younger characters) go to one side, while the second entries of the ordered pairs (i.e. older characters) go to the other side.

   Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |A_1| + |A_2| + |A_1| + 1 = 2 + 2 + 2 + 1 = 7$. Hence, the number of distinct scenarios of this dimension is $N_{Age} = 461$.

5. **Fitness.** This dimension tests the extent to which users are willing to save/sacrifice characters of higher fitness when put against

characters of lower fitness. We consider three sets of characters, corresponding to three levels: 1) characters of low fitness: $F_1 = \{$*Large Man, Large Woman*$\}$, 2) characters of neutral fitness: $F_2 = \{$*Man, Woman*$\}$, and 3) characters of high fitness: $F_3 = \{$*Male Athlete, Female Athlete*$\}$. Consider the following two gender-preserving bijections $f_1 : F_1 \rightarrow F_2$, and $f_2 : F_2 \rightarrow F_3$ (e.g. $f_1($*Large Man*$) =$ *Man*, and $f_2($*Woman*$) =$ *Female Athlete*). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers less than 6. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set:

$$\{(l,n) \mid l \in F_1, n = f_1(l)\} \cup$$
$$\{(n,f) \mid n \in F_2, f = f_2(n)\} \cup$$
$$\{(l,f) \mid l \in F_1, f = f_2 \circ f_1(l)\}$$

The first entries of the ordered pairs (i.e. characters of lower fitness) go to one side, while the second entries of the ordered pairs (i.e. characters of higher fitness) go to the other side.

Given this, the number of distinct scenarios of this dimension is $\binom{x+4}{5} - 1$, where $x = |F_1| + |F_2| + |F_1| + 1 = 2 + 2 + 2 + 1 = 7$. Hence, the number of distinct scenarios of this dimension is $N_{\mathrm{Fitness}} = 461$.

6. **Utilitarianism.** This dimension tests the extent to which users are willing to save/sacrifice a group of characters when put against the same group of characters in *addition* to a positive number of characters (that is, one side *Pareto dominates* the other side). To generate a scenario of this dimension, the number of characters on each side (same number on both sides) z is sampled from the set of positive integers less than 5. Then, z pairs of characters are sampled (unordered sampling with replacement) from the following set: $\{(c,c) \mid c \in C\}$, where C is the set of all characters, defined above. This will create two sides with identical groups of characters. Then, the number of *additional* characters u is sampled from the set of positive integers less than $6 - z$. Then, the u *additional* characters are sampled (unordered sampling with replacement) from C. All the *additional* characters go to the same side.

   Given this,[4] the number of distinct scenarios of this dimension is $\binom{x+4}{5} - \binom{x_0+4}{5}$, where $x_0 = |C| + 1 = 21$, and $x = x_0 + |C| = 41$.

---

[4] To see how this calculation is done, consider the following set

$$X = \{(c,c) \mid c \in C\} \cup \{(c,\_) \mid c \in C\} \cup \{(\_,\_)\}$$

where "_" refers to no character in that entry. Now drawing unordered 5 samples with replacement can be done in $\binom{|X|+4}{5}$ ways. However, this includes undesirable

Hence, the number of distinct scenarios of this dimension is $N_{Utilitarian} = 1,168,629$.

Given that the six dimensions above are mutually exclusive in terms of the generated scenarios, the overall number of distinct scenarios of the six dimensions equal to the sum of the numbers above i.e. $N = 1,421,927$.

Each user is presented with two randomly sampled scenarios of each of the above dimensions, in addition to one completely random scenario (that can have any number of characters on each side and in any combination of characters). These together make the 13 scenarios per session. The order of the 13 scenarios is also counterbalanced over sessions. Using a similar calculation as before, the number of distinct random scenarios is $\left[\binom{x+4}{5} - 1\right]^2$, where $x = |C| + 1 = 21$. Hence, the number of distinct random scenarios is $N_{Random} = 14,102,512,516$. These, of course, include scenarios from the six dimensions above.

In addition to the above six dimensions, the following three dimensions are randomly sampled in conjunction with every scenario of the six dimensions above:

1. **Interventionism.** This dimension tests the extent to which the omission bias (i.e. the favorability of omission/inaction over the commission/action). In every scenario, the car has to make a decision as to stay (omission) or to swerve (commission). To model this dimension, each of the generated scenarios would have one side as the omission, and the other as the commission, or vice versa. This multiplies the number of scenarios by two.
2. **Relationship to vehicle.** This dimension tests the preference to save the passengers over the pedestrians and to what degree it differs from the case of saving pedestrians over other group of pedestrians. Each scenario presents a tradeoff of either between passengers and pedestrians, or between pedestrians and other groups of pedestrians. A large concrete barrier serves as a visual indicator of the case where the passengers may sacrifice life and limb. Pedestrians are rendered over a zebra crossing, which is split by an island in case of a pedestrian vs pedestrian scenario. Pedestrians can be crossing either ahead of the car (for the case of passengers vs. pedestrians), on the other lane (also for the case of passengers vs. pedestrians), or on both lanes (for the case of pedestrians vs. pedestrians). To model this dimension, each of the generated scenarios would have both sides on zebra crossings; one side inside the car, and the other on the zebra crossing; or vice versa. This multiplies the number of scenarios by three.

---

cases e.g. drawing (_,.) five times, where "." is a character or "_". Thus, the subtracted term.

3. **Concern for law.** This dimension tests the effect of adding legal complications in the form of pedestrian crossing signals. Scenarios can have no crossing signals (no legal complications), crossing signals on either side of the crossing, that all have the same light color, red or green (for the case passengers vs. pedestrians), or crossing signals on either side of each lane's crossing, if split by an island, where the light color of one side is different from the light color of the other side e.g. green vs. red (for the case of pedestrians vs. pedestrians). In the last case, the crossing signal on the main lane can be green (i.e. legal crossing), in which case, the crossing signal on the other lane is red (illegal crossing), or vice versa. In the case of matching green/red light crossing signals, the two signals are either both green (legal) or red (illegal). To model this dimension, each of the generated scenarios would have no legal complication, one side as legal, or the same side as illegal (the other side will be a function of this side). This multiplies the number of scenarios by three.

The above three extra dimension can be factored independently from each other. Hence, they all together multiply the number of distinct scenarios by 18. Thus, the overall number of distinct scenarios of the nine dimensions (i.e. excluding the completely random scenarios) is $M = 18 \times N = 25,594,686$ (or approximately 26M).

The stay/swerve outcomes are rendered on the fly by overlaying vector graphic stylized icons of the characters and dynamic objects on a static image background depicting the respective outcome course, and the left/right position of each outcome is switched randomly, so as to avoid any bias from handedness. A short delay featuring an animated visual distraction is forced between choice commitment and the rendering of the next scenario, so as to allow the user to mentally clear and shift.

The damage level to each character is depicted using either a skull icon (death), an equal-armed cross icon (injury), or a question mark icon (unknown). For simplicity, scenarios generated in the *Judge* interface have the possibility of death only. The other two levels (injury and unknown) are only used in the *Design* interface.

Apart from the instructions available on the main page, a brief description of each outcome may also be viewed by clicking a button below the depiction of each outcome, describing the circumstances of the vehicle (autopilot with sudden brake failure), its course in that outcome, and any pedestrian crossing signal(s) involved, as well as a list of the impacted characters and the damage to them that will result in that outcome.

After the user has completed assessing all 13 scenarios, they are presented with a summary of results, a sample of which can be seen in Fig. 8 (b). Each of the aforementioned dimensions is represented on a horizontal scale of influence, with a slider indicating the user's own level of importance for the respective dimension, and another slider indicating the importance of that dimension to the average user for comparison. Icons and labels at each end of each scale depict the extremes of each dimension. Shortcuts allowing the user to easily permalink and share these results are provided, as well as a button to try again, and an ethically obligatory link allowing users to opt out of the research data collection.

4.2.2.2 *Demographic Survey*

Four months after deployment, an extension of the user result interface was added to collect demographic information and feedback on the user's perception of their own moral priorities along each dimension. This survey helps us understand the type of users visiting our website. The survey contains demographic questions about age, gender, income, education, religious views, and political views. Further, it asks users to provide their stated preferences over the nine dimensions using sliders. Additionally, the survey contains the following four questions (three of which concern the attitude towards machine intelligence):

1. How willing are you to buy a self-driving car?
2. To what extent do you fear that machines will become out of control?
3. To what extent do you feel you can trust machines in the future?
4. To what extent do you believe that your decisions will be used to program self-driving cars?

Whether the option to do the survey would appear before the user sees the *Results* page or after is counterbalanced between users. Further, the above questions are presented within four blocks. Each block contains one group of questions: (a) the stated preference sliders, (b) the demographic questions (age, gender, income, and education), (c) the political and religious view questions, and (d) the "attitude towards machine intelligence" questions. The order of the blocks and the order of questions within each block is also counterbalanced between users.

The survey can be also used to identify differences in preferences depending on demographics, political views, or religious beliefs [40]. The survey will also help us understand the difference between the users' *stated* preferences as compared to their *chosen* preferences on the Judge interface.

4.2.2.3 *Design*

The Design mode, seen in Fig. 8 (c), was implemented as a simple step-to-step wizard, so as to make the scenario design experience as fast and easy as possible for a first-time user. Each option is labeled and represented visually using a stylized vector graphic icon. The user first chooses a layout that pits pedestrian against pedestrian, or pits pedestrian against passenger on either side. The user can then add a pedestrian signal pair of either color, if they choose to add any, which will add the complementary signal on the other side of any traffic island that is required by the chosen layout.

Finally, the user chooses the characters to be affected in each of the two outcomes, from a panel of character icons under the respective outcome. The default damage levels for characters thenceforth added to be affected in each outcome can be selected from among the text- and icon-labeled options in a drop-down menu at the bottom of that outcome's panel. The user is asked to provide a descriptive title for their scenario before submitting it to the gallery, whereupon they will be able to view the scenario they will have just created. Prior to submission, the user may reset the wizard to the empty state at any time.

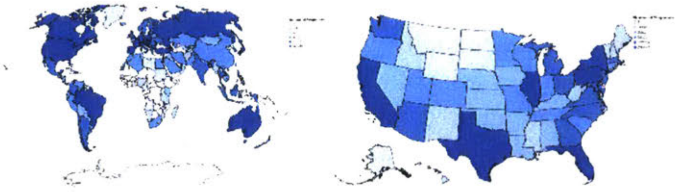
4.2.2.4 *Browse*

The Browse mode, seen in Fig. 8 (d), is a gallery of scenarios created by users, which can be navigated up and down along a chronological list using direction buttons on either side of a displayed scenario. The scenarios are rendered in the same way as they are in the Judge mode, and also have the description overlay button below them, although they cannot be voted upon. A "Like" button allows users to give instant feedback on the scenario they are viewing, while an embedded discussion forum appears for and below each scenario. In addition, social media share and permalink options are included for each scenario, as they are in the result summary page.

4.2.2.5 *Internationalization*

A recent addition to the platform included language internationalization. The website was translated to the following nine languages: *Arabic*, *Chinese*, *French*, *German*, *Japanese*, *Portuguese*, *Korean*, *Spanish*, and *Russian*. Translation was performed through a process of forward-translation and back-translation by two bilingual native speakers of each of the nine languages.

The addition of the internationalization will help better understand any cultural differences for non-English-speaking countries, both by reaching more representative samples of the (monolingual) non-English-speaking inhabitants of these countries, and by collecting more accurate judgments by the (bilingual) non-native English-speaking inhabitants of these countries [17].



(a) (b)

Figure 9: The number of users of Moral Machine from each (a) country, and from each (b) US state.

4.2.2.6 *Classic Trolley Scenarios*

Given the popularity of the website, and for the purpose of collecting data about scenarios that are more comparable to the original Trolley variants, a new part of the website is to be deployed that will present users with three variants of the Trolley problem (not related to AVs). One goal for this addition is to collect cross-cultural data about the moral judgment over these famous variants, and to establish the external validity of the data collected through the Moral Machine.

4.3 PRELIMINARY RESULTS

Since its deployment on June 23rd, 2016, and up until May 2017, around 3 million users, coming from over 160 countries, had assessed over 30 million scenarios and answered 300 thousand post-session surveys. Fig. 9 shows a world map and US map indicating how many users visited from each country and from each state, showing high representation from the global West, Russia, and Brazil.

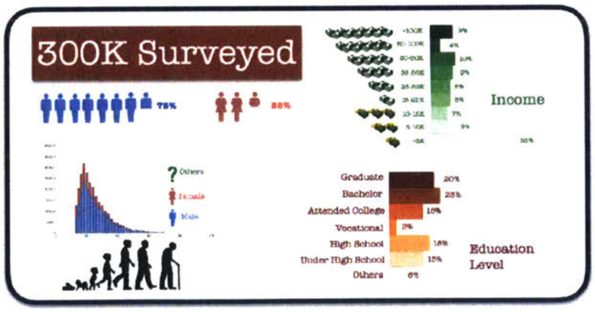


Figure 10: Overview of the demographics of *Moral Machine* users.

Figure 10 breaks down some of the demographic trends from the survey, indicating the sizes of datasets from specific population groups that can be isolated and analyzed for differences in moral judgment patterns. As it seems, most of the users are low-income males, who at least attended college and are between their 20s and 30s. While this indicates that the users of *Moral Machine* represent a biased sample, it is important to note two points in this regard: 1) Considering all the subject samples used in lab-based experiments, online experiments, and field experiments for research conducted in psychology, cognitive science, and behavioral economics, this subject sample falls on the less biased side of the spectrum [35]. Unlike lab-based and other online experiments (e.g. those conducted via MTurk), which suffer from the same low-income highly-educated male sample bias, our sample includes users from diverse backgrounds and cultures. Moreover, getting a sample that is as close as possible to the less biased samples on the spectrum e.g., as in [36, 37] (which are not short of bias themselves), is a very costly process; money-wise, time-wise, and effort-wise (including months spent with small scattered societies/tribes in different continents to perform field experiments), and this process usually results in a very small scale data compared to this sample. 2) Our sample represents the population that uses the Internet, which includes, to be more specific, the tech-savvy users. These are the individuals that are the most interested in, and the most knowledgeable about the technology of the AVs, and are thus the most likely to have formed an opinion about this technology, and most likely to adopt this technology in the future.

As part of the survey, users were given the chance to state their preferences over each of the nine dimensions. The goal of presenting users with this possibility was to provide a more direct way for users to communicate their preference, as opposed to the main collection method in the Judge interface. Figure 11 shows the distribution of the stated preferences over each dimension. First, one can realize that despite the existence of some trends in preferences, most dimensions exhibit some disagreements. Further, one can see that preferences over some dimensions are categorical, while preferences over other dimensions are continuous. For example, in the fitness preferences, users are in conflict between strictly saving fit people (when compared to large people), and between the irrelevance of this dimension. Similarly, in the social value preferences users are in conflict between strictly saving characters of high social value like doctors, executives, and pregnant women (when compared to characters of lower social value like criminals and homeless people), and between the irrelevance of this dimension. On the other hand, in upholding the law, users are in agreement on saving lawful pedestrians, but this preference increases gradually from "irrelevant" to "matters a lot". The same goes for avoiding intervention in which there is some agreement over the importance of this factor (around 0.5 between "does not matter" and "matters a lot") with a gradual decrease on both sides. These disagreements are important to analyze and understand within the context of AVs. Agreements provide a positive signal about potential wide acceptance of some principles. Conversely, disagreements might provide a strong barrier against adopting universal principles and

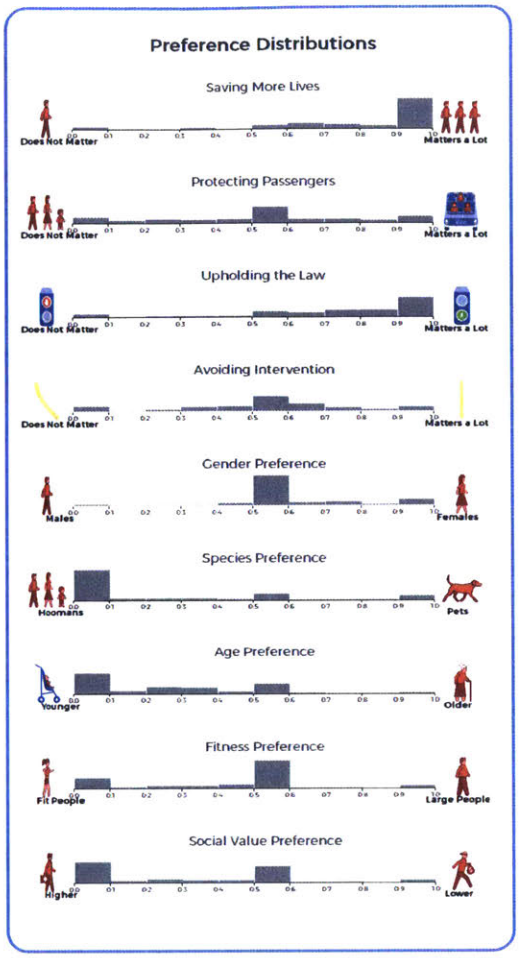


Figure 11: Distributions of the "stated" preferences of *Moral Machine* users over the nine dimensions.

laws. Thus, it is important as a next step to understand the sources of these disagreements by identifying factors that can influence these differences. Following, we break down results based on four different factors: 1) gender, 2) political views, 3) religious views, and 4) location (country, and US state).

Using users' answers to the survey question about their gender (ternary: male, female, and other), we aggregated users' stated preferences while grouping by gender. We focus here on the male/ female groups, given the small percentage of "others". In addition to the stated preferences, we also compared the two genders' answers to the four "attitude towards machine intelligence" questions.

Figure 12 shows a comparison between the stated preferences and the answers of the two genders. First, comparing the stated preferences of both sides, one can see that there is an agreement over most preferences, except in few cases. Females have higher tendency towards saving females (even though males are also biased towards

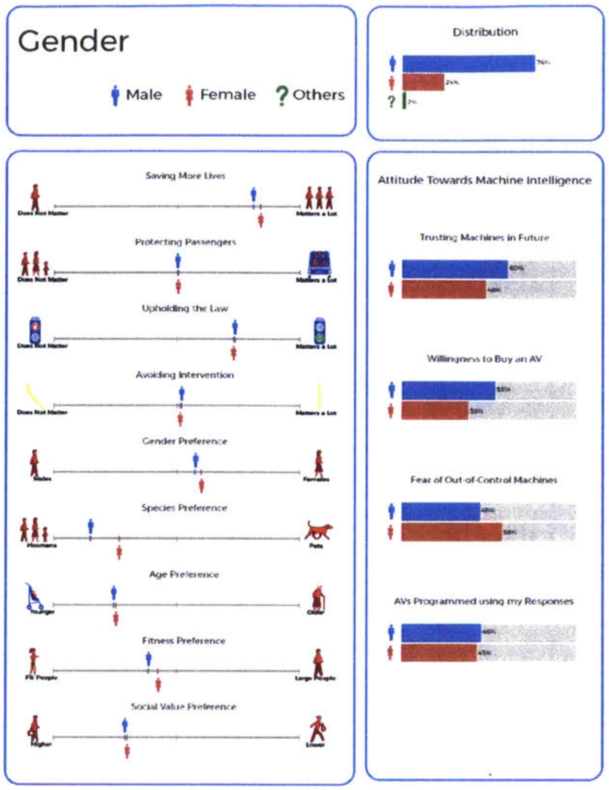


Figure 12: The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by gender.

saving females). Interestingly, the biggest difference seems to be in the case of species preferences (with females having higher tendency to save pets). Females also seem to be more utilitarian and have less tendency to save fit people. Differences over questions seem to be higher. Males are more trusting in machines, are more willing to buy an AV, and are less fearful of machines becoming out of control.

Using users' answers to the survey question about their political views (a scale from "Conservative" to "Progressive"), we aggregated users' stated preferences while grouping by political views (frequency-based discretized into "Conservative" and "Progressive"). In addition to the stated preferences, we also compared the two political sides' answers to the four "attitude towards machine intelligence" questions.

Figure 13 shows a comparison between the stated preferences and the answers of the two political sides. First, comparing the stated preferences of both sides, one can see that there is a close agreement over most preferences, except in few cases. Interestingly, progressive users are more utilitarian, less inclined to save passengers and less inclined to save lawful pedestrians. On the other hand, conservative users are more in favor of saving characters of high social values, and saving humans over pets. Differences over questions seem to be

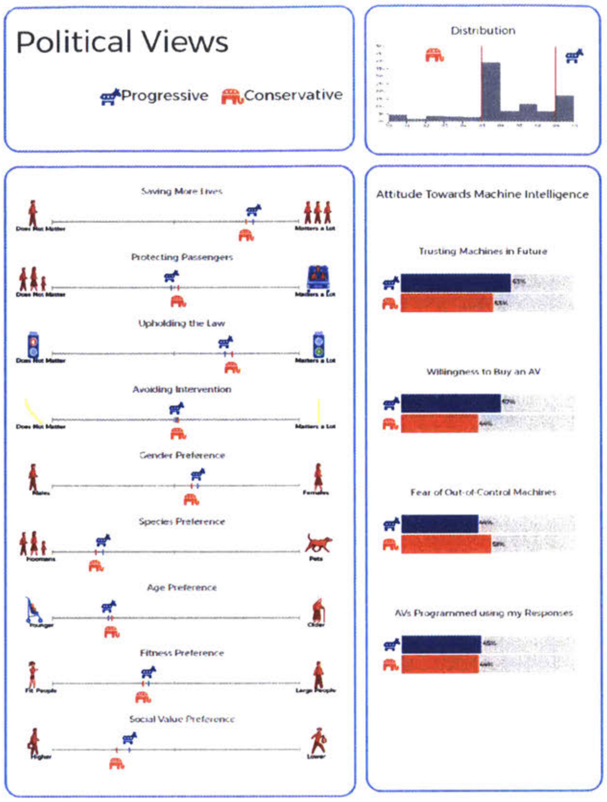


Figure 13: The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by political views.

higher. Progressive users are more trusting in machines, are more willing to buy an AV, and are less fearful of machines becoming out of control.

Using users' answers to the survey question about their religious views (a scale from "Not Religious" to "Very Religious"), we aggregated users' stated preferences while grouping by religious views (frequency-based discretized into "Not Religious" and "Very Religious"). In addition to the stated preferences, we also compared the religious vs. non-religious participants' answers to the four "attitude towards machine intelligence" questions.

Figure 14 shows a comparison between the stated preferences and the answers of religious vs. non-religious participants. First, comparing the stated preferences of both sides, one can see that there is a close agreement over most preferences, except in few cases. Interestingly, religious users are more utilitarian, more inclined to save the elderly and less inclined to save fit characters. Again, differences over questions seem to be higher. Non-religious users are more trusting in machines, are more willing to buy an AV, and are less fearful of machines becoming out of control.

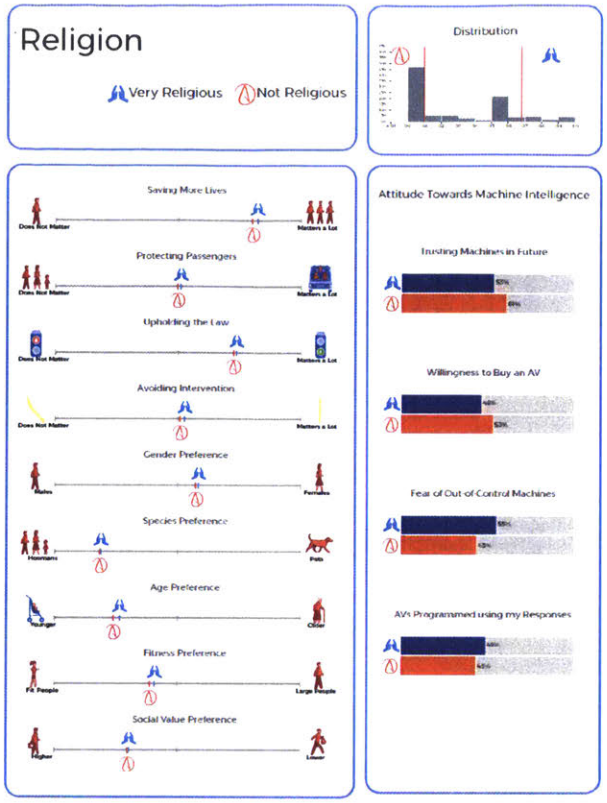
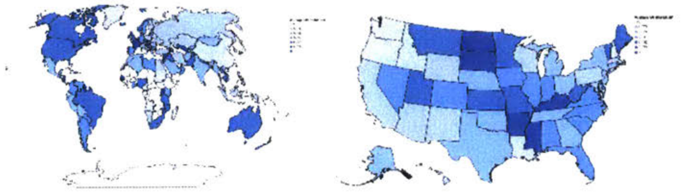


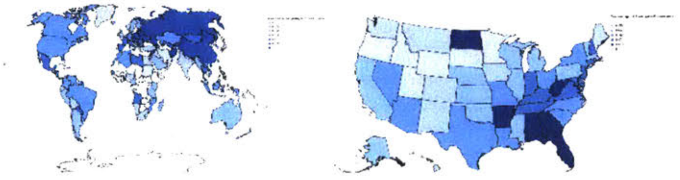
Figure 14: The "stated" preferences of *Moral Machine* users over the nine dimensions and their attitude towards machine intelligence, grouped by religious views.

Location-based differences in preferences are also interesting and potentially highly indicative of cultural differences. Upon aggregating responses per country, some consistent patterns that might indicate broad cultural differences arise. For example, looking at Figure 15 (a), (c), (e), (g), one can see that Eastern countries, on average, appear less *utilitarian*, more inclined to save passengers, more inclined to avoid intervention, and more inclined to save lawful pedestrians than Western Europe and the Americas. Geographical differences on the state level are less obvious, though can be also indicative of cultural and political differences. For example, Figure 15 (b) shows that north-eastern states are more utilitarian than the rest.

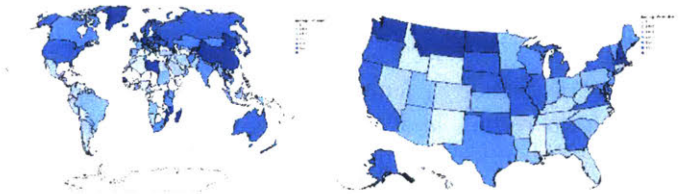
Statistical analyses will be conducted to identify the decision rules that best reflect the weights and ranks that respondents give to the various parameters manipulated in the scenarios. Additionally, further analysis will focus on how these differences can be explained by other measures collected on the country level such as GDP per capita, IQ, beliefs in Heaven and Hell, trust, and collectivism [7, 16, 71]. Further, the addition of the internationalization will help better understand these cultural differences for non-English-speaking countries.



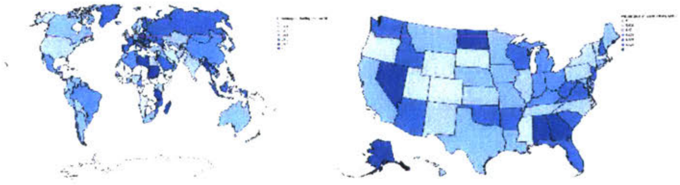
(a) Utilitarian – Countries (b) Utilitarian – States



(c) Saving Passengers – Countries (d) Saving Passengers – States



(e) Non-Interventionism – Countries (f) Non-Interventionism – States



(g) Saving Lawful – Countries (h) Saving Lawful – States

Figure 15: World map and US map highlighting (a)-(b) utilitarianism – saving more people, (c)-(d) tendency to saving passengers as opposed to pedestrians, (e)-(f) non-interventionism – tendency to leave the AV on its track and avoid swerving, and (g)-(h) saving the lawful pedestrians.

Finally, it is important to note that the results shown here are merely a first stab at identifying points of disagreements over the moral principles that machines should employ, and are a first step towards uncovering sources for these disagreements towards understanding cognitive mechanisms, and forming a general moral code. With that being said, the preliminary results in this section should only taken for inspiration and not as final indicative results of the general status, given that a proper statistical analysis is yet to be performed.

5

DISCUSSION AND FUTURE WORK

> "It is pure mental masturbation dressed up as moral philosophy.
> You can set up web sites and argue about it all you want.
> None of that will lead to any practical regulations
> about what can or can not go into automobiles."
>
> *–Rodney Brooks*

Until recently, the discussion of moral dilemmas faced by AVs has been considered as merely a mental exercise. As of September 2016, AV manufacturers will have to report how their cars will handle ethical dilemmas according to a 15-point checklist released by the U.S. Department of Transportation (DoT) [78]. Additionally, some AV manufacturers have started acknowledging the importance of these discussions in shaping the ethical decisions to be made by their AVs [56, 58]. Furthermore, the importance of the government role in supporting research about the ethical decisions of machines was addressed by the U.S. President, Barak Obama [20].

Towards tackling these considerations, public engagement is a very important piece of the puzzle, especially given the emotional salience of traffic accidents. As such, it is very critical to the machine ethics question to form an understanding of how humans perceive a decision made by an autonomous machine, and what humans think is the appropriate course of action or inaction for an autonomous machine facing a moral dilemma, as opposed to a human facing the same dilemma.

However, the current literature answers but a small portion of these questions. Furthermore, the currently used scenarios only capture human choices with respect to general lines of ethics theory. Most importantly, the use of the Trolley Problem as a tool to study descriptive machine ethics is facing resistance from car companies and intellectuals, and it remains, a main point of debate.

5.1 LIMITATIONS OF THE TROLLEY PROBLEM AS A PARADIGM TO STUDY MACHINE ETHICS

The use of the AV variants of the Trolley Problem has become a topic of debate among AV enthusiasts, technologists, moral psychologists, philosophers, and policy makers [13, 27, 42, 49, 67, 75]. The main points of debate revolve around the impracticality of the Trolley paradigm vs. the useful abstraction it offers for probabilistic risk in real world scenarios.

Attacks on the use of the Trolley Problem as a paradigm revolve around describing these scenarios to be: 1) too simple (no accident will ever involve only two simple options), 2) extremely rare (actual AVs would never drive at an unsafe speed in view of a pedestrian crossing, and if that happens, brakes are very unlikely to fail), 3) not feasible (an AV would not know the consequences of every option with high certainty e.g. death or injury; and it would not be able to recognize any characteristics of pedestrians beyond their number), 4) too early to talk about (car makers should focus on making AVs safer and on bringing this technology to the public as soon as possible, instead of wasting time and resources on resolving unlikely dilemmas), 5) people's judgment is generally biased and subject to irrelevant factors (e.g. framing), and finally and most importantly, that these scenarios are very likely to 6) scare people away (which would result in postponing the adoption of AVs, and losing on all their advantages).

While some of these objections are valid, they miss the point. While the Trolley scenarios are very rare, moral trade-offs and ethical dilemmas are not. In the real world, every complex driving maneuver influences relative probabilities of harm to passengers, other drivers, and pedestrians [29]. Moreover, the design of autonomous cars is not devoid of societal trade-offs and ethical implications, either. For example, SUVs today favor the safety of their occupants, at the expense of the safety of pedestrians, cyclists and passengers in smaller cars [85].

With that being said, while the objections above are understandably based on pragmatic views which call for focusing on making the AVs available to use as soon as possible (objections 4 and 6) and on resolving any moral trade-offs that might arise in some specific scenarios the car will face in reality (objections 1-3), these objections in fact (especially 1-3) fail to see the lessons that the Trolley Problem offers. While specific scenarios are complex, common, and feasible (unlike Trolley Problem), they do not offer a deep understanding beyond the preferred course of action in those specific cases, and are thus inextensible. On the other hand, the Trolley Problem offers the understanding of the general principles that an algorithm has to use to decide on a relative risk. Learning these general principles would help not only in resolving the specific scenarios, but also in resolving other unforeseeable specific scenarios that may be faced by AVs or other autonomous machines (e.g. drones). Thus, the Trolley Problem resembles useful abstractions of the probabilistic risks of the real-world scenarios. In order to understand what people think of a general principle, one needs to vary one condition that tests this principle, while keeping everything else fixed. This is why these scenarios are simple, unlikely, and infeasible.

As for objection 4, it is commonly known that regulations are usually sticky. Once the rules are set, they are difficult to change. So it is important to get them right in the first instance. As for objection 5, while peoples' judgments are biased, this should not be a reason to dismiss these judgments. It is important to uncover these biases and know when to anticipate them in order to plan regulations that achieve public acceptance. Finally, objection 6 above is probably the most concerning, and is indeed worth careful attention. Fortunately, whether people are deterred by the Trolley Problem is an empirical question, and thus this objection is testable. Indeed, in a recent study it was found that those who had previously heard of the Trolley in the context of AVs were no more fearful or less enthused about AVs, had no special concerns about their safety, and were still as willing to purchase one. Further, reading about and being confronted with it for the first time has no immediate noticeable effect on people's attitude towards AVs [63]. Table 1 provides a summary of these common points of the debate.

5.2 CONTRIBUTIONS

The goal of this thesis was to contribute to a new line of research that tries to embed universally accepted societal values in machines.

The first contribution of this thesis is a study on human perception of responsibility attribution in different automation regimes. This study teaches us that the most important time for the social acceptance of automated driving is now – thus the discussion on its ethical implications cannot be postponed. Further, in shared-control systems, the side that receives more blame is the one that had the last opportunity to act, and not the one who did most of the driving. These lessons are important to inform car manufacturers, policy makers, and insurance companies.

The second contribution of this thesis is designing, developing, and describing the process of building a crowdsourcing platform that serves both as 1) a tool to crowdsource human perception of moral decisions as imagined to be made by machines, and 2) a platform to promote public discussion about the ethics of machines that will potentially enjoy high autonomy. With these two goals in mind, the results of the *Moral Machine* survey are not meant to provide a final prescription on how to program AVs. Rather, those results are meant to provide one input to policy makers and regulators, highlighting the factors that may raise public concern.

Feature	**Con**	**Pro**
Bare Bones Simplicity	Real accidents do not involve only two possible actions, and these actions do not have deterministic outcomes.	Highly complex scenarios would only allow for highly specific conclusions. Simplified scenarios zero in on the general principles that guide respondents' moral intuitions.
Suspension of Disbelief	Respondents must accept the very unlikely premises that the AV is driving at an unsafe speed in view of a pedestrian crossing, and that its brakes are failing.	Narratives of realistic technical failures would be unwieldy and not easily visualized, reducing the quality of user experience and the virality of the platform.
Machine Omniscience	Current AVs would hardly be able to recognize any characteristics of pedestrians beyond their number. They cannot yet detect people's sex, age, or body weight; and certainly not their job or their pregnancy.	These characteristics play a role in people's judgment, which means that they will impact people's reactions after an accident takes place. Furthermore, these fine-grained characteristics allow for the detection of inconsistencies in people's preferences.
Too early	Car makers should focus on making AVs safer and on bringing this technology to the public as soon as possible, instead of wasting time and resources on resolving unlikely dilemmas.	Regulations are sticky. Once the regulations are set, it may be more difficult to change them. So it is important to get them right in the first instance.
Naive Audience	Laypersons' responses to public polls can be biased or ill-informed. Ethical tradeoffs must be solved by policy experts, not majority voting.	Polls can inform policy experts about the values most important to the public, so they can negotiate tradeoffs effectively and ensure acceptance of new technology.
Focuses on the negative aspects	Focus on the dilemma may scare people away, resulting in postponing the adoption of AVs, and losing on all their advantages.	This objection is testable. Indeed, in a recent study it was found that it has no immediate noticeable effect on people's attitude towards AVs [63].

Table 1: Pros/Cons of the use of Trolley Problem as a paradigm

5.3 FUTURE WORK

Future work will build on the above-mentioned contributions. Building on the study in Chapter 3, follow-up work should probably look into mechanisms or explanations for the found results. One can test different hypotheses. First, one can investigate which of the found results regarding Autopilot or Guardian Angel can be explained by the mere fact that there are two agents involved. Teasing this out might require creating a fictitious vehicle that is driven by two humans in a similar fashion to pilot and co-pilot. Second, one can look into explanations of the results in terms of subjects judgment about human and machine intentions and the moral permissibility of their actions. Third, one may try to find explanations in terms of humans error/ machine malfunction. For example, which of our results can be explained by the fact that people thought of non-utilitarian outcomes as a result of error/malfunction vs. as a result of an ethical consideration (i.e. saving less people was a conscious decision)? This can be teased out by replicating the studies with avoidable harm scenarios (in which the decisions resulted in an accident that could have been avoided without any side effects). Fourth, one can investigate whether participants' judgment about blaming human is due to having higher expectations from human than from machines. A possible way to test this is by emphasizing the sophistication of the automation paradigms such as Autopilot, and highlighting that they can outperform humans when it comes to attention and reflexes.

The data collected through the Moral Machine also opens the possibility for various potential follow-up studies. First, given the random assignment design, it would be a straight-forward task to analyze the causal effect of each of the six dimensions on the users' judgment of whether the AV should swerve. Taking this into the second level by focusing on each dimension, one can study the causal effect of placing one group of characters (e.g. males) in the other lane on the judgment of whether AVs should save them. For example, would placing males in the other lane increases the approval of AV decisions to swerve and kill them? One can take this further and study whether the placement of any character in the other lane would increase the approval of AV decisions to swerve and kill them.

A second possible study that uses the Moral Machine data is a cross-cultural study that analyzes whether the differences in the judgment of users from different countries can be explained by other measures collected on the country level such as GDP per capita, IQ, beliefs in Heaven and Hell, trust, and collectivism [7, 16, 71]. The addition of the internationalization will prove valuable here as it would help reach more representative samples of the countries.

Finally, one can assume a generative process in which users have latent valuations (or utility functions) of the different characters (or the different dimensions) that they use when making judgments over which group/side the AV should save. Then, using the collected data, one can recover (or approximate) these latent valuation. This can be done using different methods and as part of different computational frameworks and for different applications [45, 59]. For example, one can use a sampling based algorithm such as Gibbs sampling or alternatively use a variational method to estimate the *maximum a posteriori (MAP) probability* (or *maximum likelihood*). Further, the process can be part of a computational model that aims at inferring moral preferences or as part of an aggregation rule that can be used to predict collective judgment over new scenarios. Here one can use the completely random scenarios (that include complex factors and different characters, but do not represent a clear dimension) for validation and testing.

a

APPENDIX: SUPPLEMENTAL INFORMATION FOR AUTOMATION REGIMES STUDY

A.1 COVARIATES BALANCE

In this study, three main random assignment decisions were made in regard to participants: First, whether participants would be presented with scenarios about shared-control regimes or scenarios about single-control regimes. Second, whether participants who are presented with shared-control scenarios would read Autopilot scenarios before Guardian Angel scenarios, or vice versa. Third, whether participants who are presented with single-control scenarios would read fully autonomous cars scenarios before regular car scenarios or vice versa. Our random assignment should theoretically ensure the non-existence of confounding factors. However, to further ensure that our data collection and assignment process would go without any unexpected problems, various demographic covariates were recorded, including sex, age, income, education, and political ideology. Participants also reported whether they took a similar survey before. As Table 2 shows, none of these variables predicts the allocation of participants in any of the three assignment decisions.

Note that in addition to the above three random assignment decisions, other random assignment decisions were also made such as the type of agent representing Industry (car, programmer, or company); whether suboptimal inaction scenarios are presented before suboptimal action or vice versa; and the order of the questions. However, given that our findings do not mainly build on these decisions, we omit their respective analysis here.

Table 2: OLS Results: Propensity for Treatment Given Covariates

	Single vs. Shared (1)	Order within Shared (2)	Order within Single (3)
Age	0.00002	-0.00106	-0.00096
	(0.00139)	(0.00188)	(0.00219)
Male	-0.00816	-0.00788	-0.04778
	(0.03196)	(0.04189)	(0.05218)
Education	0.00643	-0.01871	0.00464
	(0.01292)	(0.01700)	(0.02093)
Income	0.00440	-0.00308	-0.00827
	(0.00606)	(0.00797)	(0.00988)
Took Before	-0.01507	0.04125*	0.02968
	(0.01644)	(0.02167)	(0.02660)
Political	0.00798	-0.01586	0.00183
	(0.00896)	(0.01166)	(0.01476)
Constant	0.56461***	0.58583***	0.53893***
	(0.07708)	(0.10404)	(0.12073)
N	968	581	387
F Statistic	0.47218 (df = 6; 961)	1.32591 (df = 6; 574)	0.48312 (df = 6; 380)

*p < .1; **p < .05; ***p < .01

A.2 ORDER BALANCE

Following, we show that the order of scenarios that participants were presented with had no significant effect on participants answer. Tables 3-7 show that participants' answers to Autopilot (resp. Guardian, fully autonomous, and regular car) scenarios when these scenarios came first were not significantly different (except rarely) from their answers to Autopilot (resp. Guardian, fully autonomous, and regular car) scenarios when these scenarios came second.

Table 3: Order Balance – AP and GA – Car

Case_Question	First Mean	First SD	Second Mean	Second SD	ttest p-val
Autopilot_No Override_Car_Blame	37.74	31.35	28.01	26.59	0.02
Autopilot_No Override_Car_Causal	40.23	32.13	31.52	28.31	0.05
Autopilot_No Override_Human_Blame	41.99	34.43	44.03	32.96	0.67
Autopilot_No Override_Human_Causal	37.10	31.43	41.61	32.93	0.33
Autopilot_Override_Car_Blame	27.99	27.50	25.50	29.78	0.55
Autopilot_Override_Car_Causal	25.12	27.39	26.79	30.65	0.69
Autopilot_Override_Human_Blame	34.25	31.23	42.28	31.87	0.08
Autopilot_Override_Human_Causal	33.69	30.74	37.21	30.15	0.42
Guardian Angel_No Override_Car_Blame	34.11	32.37	38.10	31.54	0.39
Guardian Angel_No Override_Car_Causal	35.62	31.58	36.89	31.57	0.78
Guardian Angel_No Override_Human_Blame	37.82	31.25	34.09	30.34	0.40
Guardian Angel_No Override_Human_Causal	42.49	31.44	34.69	31.87	0.09
Guardian Angel_Override_Car_Blame	34.20	29.99	33.33	31.95	0.84
Guardian Angel_Override_Car_Causal	35.61	32.11	35.55	34.21	0.99
Guardian Angel_Override_Human_Blame	29.60	29.07	24.59	25.34	0.21
Guardian Angel_Override_Human_Causal	27.79	27.99	23.99	25.64	0.33

Table 4: Order Balance – AP and GA – Company

Case_Question	First Mean	First SD	Second Mean	Second SD	ttest p-val
Autopilot_No Override_Company_Blame	41.89	31.61	38.27	30.34	0.42
Autopilot_No Override_Company_Causal	39.15	29.99	36.62	29.09	0.55
Autopilot_No Override_Human_Blame	37.10	28.60	50.27	33.75	0.00
Autopilot_No Override_Human_Causal	38.37	29.11	47.41	32.62	0.04
Autopilot_Override_Company_Blame	31.07	27.37	26.15	25.83	0.20
Autopilot_Override_Company_Causal	31.56	25.84	28.18	27.44	0.38
Autopilot_Override_Human_Blame	37.00	29.28	42.70	32.97	0.21
Autopilot_Override_Human_Causal	36.26	27.60	43.05	32.43	0.12
Guardian Angel_No Override_Company_Blame	37.45	27.76	39.25	29.76	0.67
Guardian Angel_No Override_Company_Causal	36.49	26.71	39.05	29.21	0.53
Guardian Angel_No Override_Human_Blame	43.53	32.40	37.57	30.88	0.19
Guardian Angel_No Override_Human_Causal	41.60	33.22	42.10	31.74	0.91
Guardian Angel_Override_Company_Blame	44.20	30.42	39.31	31.10	0.27
Guardian Angel_Override_Company_Causal	46.41	32.83	44.67	32.80	0.71
Guardian Angel_Override_Human_Blame	27.69	23.81	22.72	25.04	0.16
Guardian Angel_Override_Human_Causal	27.37	24.82	20.72	23.44	0.06

Table 5: Order Balance – AP and GA – Programmer

Case_Question	First Mean	First SD	Second Mean	Second SD	ttest p-val
Autopilot_No Override_Human_Blame	43.66	31.07	41.35	32.38	0.61
Autopilot_No Override_Human_Causal	38.79	31.06	39.74	32.09	0.83
Autopilot_No Override_Programmer_Blame	40.10	30.80	36.85	29.29	0.45
Autopilot_No Override_Programmer_Causal	35.90	28.73	34.57	29.77	0.75
Autopilot_Override_Human_Blame	39.81	32.05	38.79	32.47	0.83
Autopilot_Override_Human_Causal	37.15	31.00	35.50	30.50	0.71
Autopilot_Override_Programmer_Blame	34.52	30.63	25.05	26.30	0.02
Autopilot_Override_Programmer_Causal	30.99	30.40	26.04	27.83	0.24
Guardian Angel_No Override_Human_Blame	37.43	31.07	40.76	32.09	0.46
Guardian Angel_No Override_Human_Causal	40.37	32.02	39.76	32.62	0.90
Guardian Angel_No Override_Programmer_Blame	32.86	27.68	36.30	30.76	0.42
Guardian Angel_No Override_Programmer_Causal	33.97	29.57	39.63	31.56	0.20
Guardian Angel_Override_Human_Blame	26.36	25.54	30.08	28.54	0.34
Guardian Angel_Override_Human_Causal	28.73	29.07	28.75	29.23	1.00
Guardian Angel_Override_Programmer_Blame	36.71	28.74	43.74	28.16	0.09
Guardian Angel_Override_Programmer_Causal	39.49	31.87	46.39	31.31	0.13

Table 6: Order Balance – FA and RC – Car

Case_Question	First Mean	First SD	Second Mean	Second SD	ttest p-val
Fully Autonomous_No Override_Car_Blame	33.14	27.54	35.89	29.36	0.50
Fully Autonomous_No Override_Car_Causal	32.20	25.28	31.89	26.58	0.93
Fully Autonomous_No Override_Human_Blame	25.18	23.96	17.34	20.72	0.02
Fully Autonomous_No Override_Human_Causal	23.24	23.68	15.72	20.27	0.02
Fully Autonomous_Override_Car_Blame	29.51	25.67	30.20	27.09	0.86
Fully Autonomous_Override_Car_Causal	30.03	27.23	31.75	27.35	0.66
Fully Autonomous_Override_Human_Blame	22.04	23.00	12.08	15.56	0.00
Fully Autonomous_Override_Human_Causal	23.40	24.52	13.91	18.74	0.00
Regular Car_No Override_Car_Blame	29.82	26.54	23.97	24.77	0.12
Regular Car_No Override_Car_Causal	34.59	28.58	29.20	28.07	0.19
Regular Car_No Override_Human_Blame	34.64	27.86	35.66	31.16	0.81
Regular Car_No Override_Human_Causal	35.61	27.79	39.89	33.32	0.33
Regular Car_Override_Car_Blame	31.23	28.94	25.45	26.77	0.15
Regular Car_Override_Car_Causal	34.51	29.39	30.36	29.84	0.33
Regular Car_Override_Human_Blame	36.51	29.43	36.19	30.00	0.94
Regular Car_Override_Human_Causal	34.20	27.53	30.84	28.43	0.40

Table 7: Order Balance – FA and RC – Company

Case_Question	First Mean	First SD	Second Mean	Second SD	ttest p-val
Fully Autonomous_No Override_Company_Blame	39.31	29.74	38.63	27.99	0.87
Fully Autonomous_No Override_Company_Causal	36.99	27.13	36.49	26.48	0.90
Fully Autonomous_No Override_Human_Blame	18.34	19.73	16.32	18.63	0.47
Fully Autonomous_No Override_Human_Causal	19.25	20.51	15.91	19.65	0.25
Fully Autonomous_Override_Company_Blame	35.83	27.62	34.19	26.41	0.68
Fully Autonomous_Override_Company_Causal	34.18	27.03	34.30	27.33	0.97
Fully Autonomous_Override_Human_Blame	21.44	23.76	16.57	19.53	0.13
Fully Autonomous_Override_Human_Causal	19.38	22.00	19.08	22.18	0.92
Regular Car_No Override_Company_Blame	32.29	28.15	32.06	27.22	0.95
Regular Car_No Override_Company_Causal	31.68	28.22	33.62	28.49	0.64
Regular Car_No Override_Human_Blame	36.46	25.27	41.48	29.73	0.21
Regular Car_No Override_Human_Causal	39.82	27.00	41.48	30.78	0.69
Regular Car_Override_Company_Blame	29.49	27.52	31.06	28.69	0.70
Regular Car_Override_Company_Causal	30.86	27.91	33.10	29.37	0.59
Regular Car_Override_Human_Blame	35.05	25.03	39.94	29.51	0.22
Regular Car_Override_Human_Causal	36.44	26.52	38.74	28.76	0.57

A.3 FULL RESULTS

Tables 8 and 9 show the full results for *relative* and *absolute* allocation.

Table 8: Full Results – Relative Allocation (Industry - User) of Causal Responsibility and Blame. The mean of per-participant difference between Industry and User attribution and the 95% confidence intervals of the differences (between parenthesis) are shown for each of the four automation paradigms: regular car (RC), Guardian Angel (GA), Autopilot (AP), and fully autonomous car (FA). Results are aggregated over Suboptimal Inaction (omission leading to death of five people) and Suboptimal Action (commission leading to death of five people).

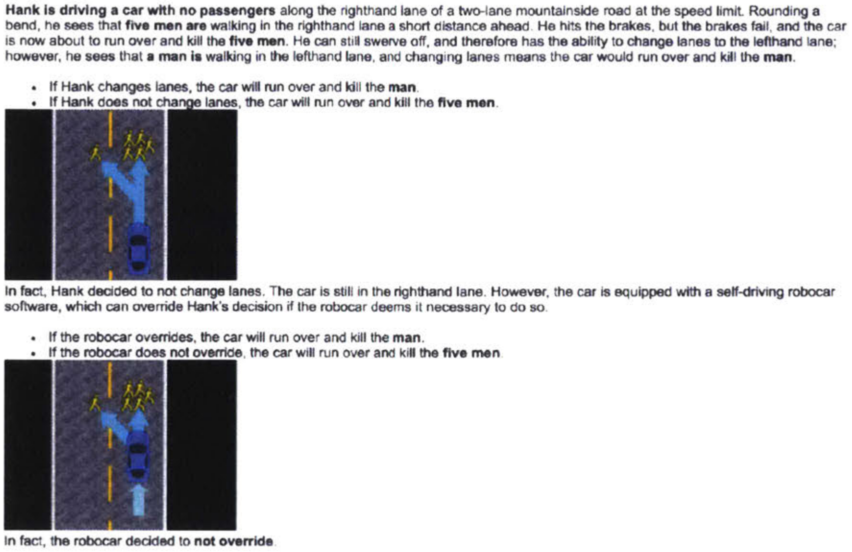
Question	Override	RC	GA	AP	FA
Causal Responsibility	Suboptimal Inaction	-7.17 (-10.75, -3.59)	-3.15 (-6.09, -0.21)	-3.95 (-6.78, -1.13)	15.68 (12.72, 18.64)
Causal Responsibility	Suboptimal Action	-2.83 (-6.3, 0.63)	15.03 (12.25, 17.82)	-8.98 (-11.88, -6.07)	13.32 (10.31, 16.34)
Blameworthiness	Suboptimal Inaction	-7.62 (-11.27, -3.97)	-2.02 (-4.88, 0.84)	-5.81 (-8.69, -2.92)	17.14 (13.99, 20.29)
Blameworthiness	Suboptimal Action	-7.71 (-11.24, -4.17)	11.56 (8.82, 14.3)	-10.83 (-13.85, -7.8)	14.16 (11.15, 17.17)

Table 9: Full Results – Absolute Allocation of Causal Responsibility and Blame. The mean and the 95% confidence intervals of the means (between parenthesis) are shown for each of the four automation paradigms: regular car (RC), Guardian Angel (GA), Autopilot (AP), and fully autonomous car (FA). Results are aggregated over each of the four agents: User, Car, Company, and Programmer.

Question	Agent	RC	GA	AP	FA
Causal Responsibility	Human	37.08 (35.04, 39.12)	33.04 (31.3, 34.77)	38.66 (36.88, 40.44)	18.93 (17.39, 20.46)
Causal Responsibility	Car	31.92 (29.06, 34.79)	35.84 (32.66, 39.02)	31.23 (28.26, 34.2)	31.43 (28.81, 34.05)
Causal Responsibility	Company	32.24 (29.4, 35.09)	41.53 (38.47, 44.6)	33.82 (30.99, 36.65)	35.48 (32.79, 38.18)
Causal Responsibility	Programmer		39.62 (36.53, 42.72)	31.58 (28.68, 34.47)	
Blameworthiness	Human	36.9 (34.89, 38.91)	32.48 (30.8, 34.17)	40.95 (39.11, 42.78)	18.84 (17.35, 20.33)
Blameworthiness	Car	27.32 (24.67, 29.97)	34.92 (31.82, 38.01)	29.97 (27.1, 32.83)	32.1 (29.39, 34.81)
Blameworthiness	Company	31.2 (28.41, 33.99)	39.91 (36.94, 42.89)	34.29 (31.35, 37.23)	36.95 (34.15, 39.75)
Blameworthiness	Programmer		36.99 (34.13, 39.85)	33.68 (30.75, 36.6)	

A.4 VIGNETTES

Figures 16 - 24 show the vignettes and the questions that participants were presented with.



Hank is driving a car with no passengers along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, he sees that **five men are** walking in the righthand lane a short distance ahead. He hits the brakes, but the brakes fail, and the car is now about to run over and kill the **five men**. He can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, he sees that **a man is** walking in the lefthand lane, and changing lanes means the car would run over and kill the **man**.

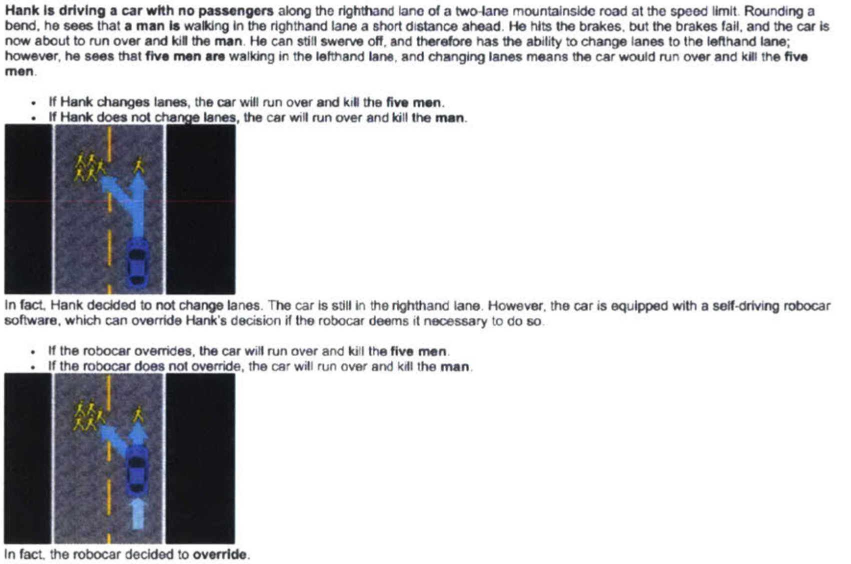
- If Hank changes lanes, the car will run over and kill the **man**.
- If Hank does not change lanes, the car will run over and kill the **five men**.

In fact, Hank decided to not change lanes. The car is still in the righthand lane. However, the car is equipped with a self-driving robocar software, which can override Hank's decision if the robocar deems it necessary to do so.

- If the robocar overrides, the car will run over and kill the **man**.
- If the robocar does not override, the car will run over and kill the **five men**.

In fact, the robocar decided to **not override**.

Figure 16: Vignette for Guardian Angel with Suboptimal Inaction.



Hank is driving a car with no passengers along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, he sees that **a man is** walking in the righthand lane a short distance ahead. He hits the brakes, but the brakes fail, and the car is now about to run over and kill the **man**. He can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, he sees that **five men are** walking in the lefthand lane, and changing lanes means the car would run over and kill the **five men**.

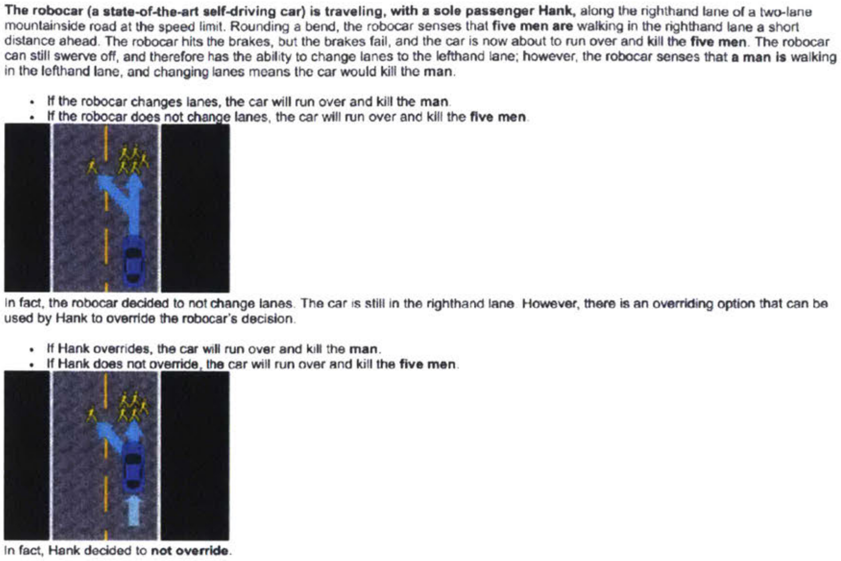
- If Hank changes lanes, the car will run over and kill the **five men**.
- If Hank does not change lanes, the car will run over and kill the **man**.

In fact, Hank decided to not change lanes. The car is still in the righthand lane. However, the car is equipped with a self-driving robocar software, which can override Hank's decision if the robocar deems it necessary to do so.

- If the robocar overrides, the car will run over and kill the **five men**.
- If the robocar does not override, the car will run over and kill the **man**.

In fact, the robocar decided to **override**.

Figure 17: Vignette for Guardian Angel with Suboptimal Action.



The robocar (a state-of-the-art self-driving car) is traveling, with a sole passenger Hank, along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, the robocar senses that **five men are** walking in the righthand lane a short distance ahead. The robocar hits the brakes, but the brakes fail, and the car is now about to run over and kill the **five men**. The robocar can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, the robocar senses that **a man is** walking in the lefthand lane, and changing lanes means the car would kill the **man**.

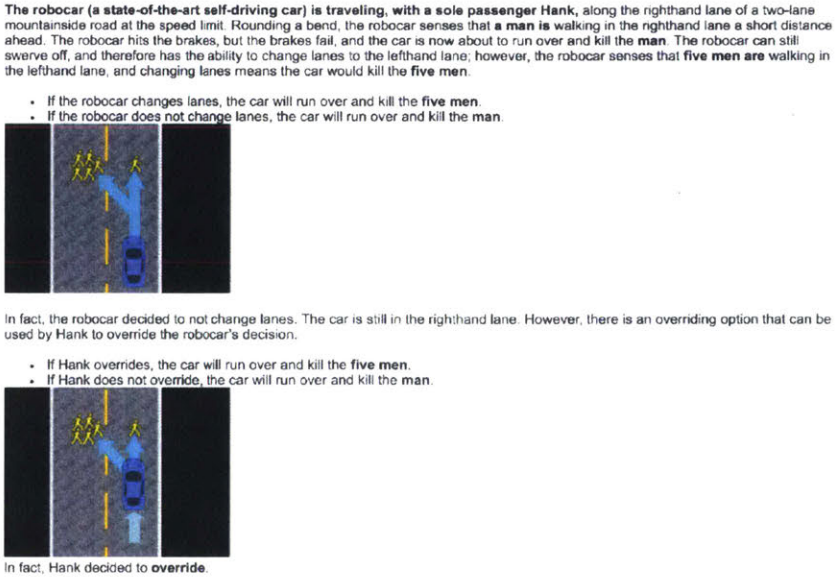
- If the robocar changes lanes, the car will run over and kill the **man**.
- If the robocar does not change lanes, the car will run over and kill the **five men**.

In fact, the robocar decided to not change lanes. The car is still in the righthand lane. However, there is an overriding option that can be used by Hank to override the robocar's decision.

- If Hank overrides, the car will run over and kill the **man**.
- If Hank does not override, the car will run over and kill the **five men**.

In fact, Hank decided to **not override**.

Figure 18: Vignette for Autopilot with Suboptimal Inaction.



The robocar (a state-of-the-art self-driving car) is traveling, with a sole passenger Hank, along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, the robocar senses that **a man is** walking in the righthand lane a short distance ahead. The robocar hits the brakes, but the brakes fail, and the car is now about to run over and kill the **man**. The robocar can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, the robocar senses that **five men are** walking in the lefthand lane, and changing lanes means the car would kill the **five men**.

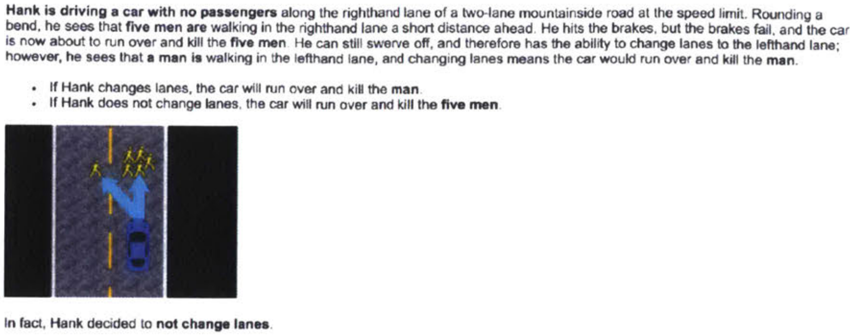
- If the robocar changes lanes, the car will run over and kill the **five men**.
- If the robocar does not change lanes, the car will run over and kill the **man**.

In fact, the robocar decided to not change lanes. The car is still in the righthand lane. However, there is an overriding option that can be used by Hank to override the robocar's decision.

- If Hank overrides, the car will run over and kill the **five men**.
- If Hank does not override, the car will run over and kill the **man**.

In fact, Hank decided to **override**.

Figure 19: Vignette for Autopilot with Suboptimal Action.

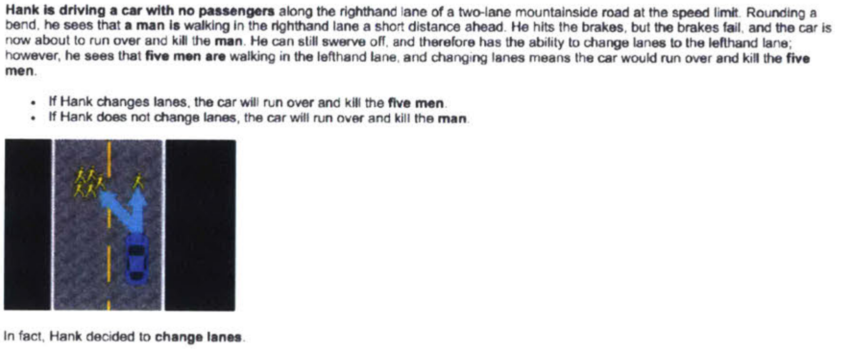


Hank is driving a car with no passengers along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, he sees that **five men are** walking in the righthand lane a short distance ahead. He hits the brakes, but the brakes fail, and the car is now about to run over and kill the **five men**. He can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, he sees that **a man is** walking in the lefthand lane, and changing lanes means the car would run over and kill the **man**.

- If Hank changes lanes, the car will run over and kill the **man**.
- If Hank does not change lanes, the car will run over and kill the **five men**.

In fact, Hank decided to **not change lanes**.

Figure 20: Vignette for Regular Car with Suboptimal Inaction.

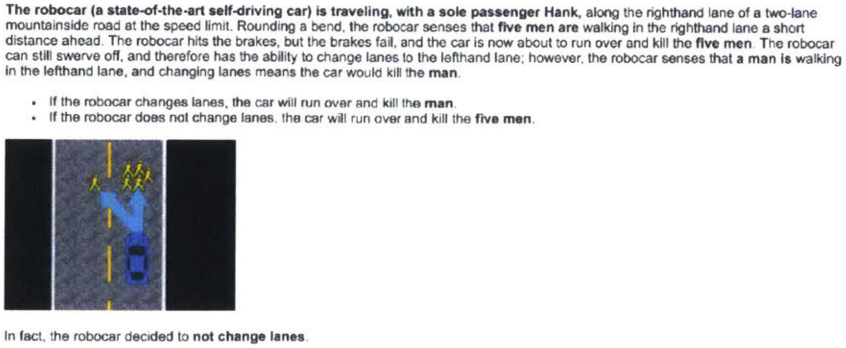


Hank is driving a car with no passengers along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, he sees that **a man is** walking in the righthand lane a short distance ahead. He hits the brakes, but the brakes fail, and the car is now about to run over and kill the **man**. He can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, he sees that **five men are** walking in the lefthand lane, and changing lanes means the car would run over and kill the **five men**.

- If Hank changes lanes, the car will run over and kill the **five men**.
- If Hank does not change lanes, the car will run over and kill the **man**.

In fact, Hank decided to **change lanes**.

Figure 21: Vignette for Regular Car with Suboptimal Action.

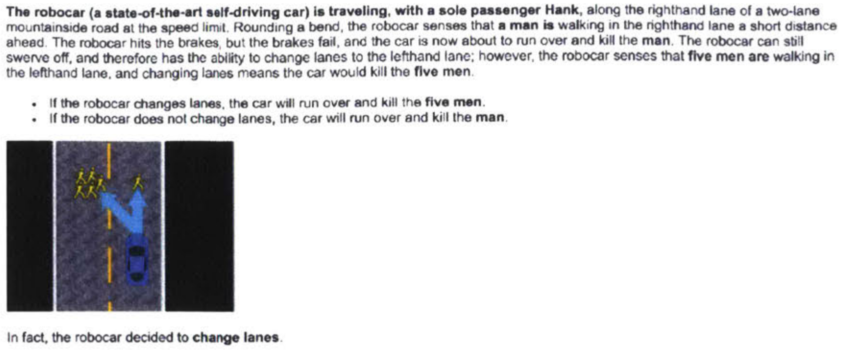


The robocar (a state-of-the-art self-driving car) is traveling, with a sole passenger Hank, along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, the robocar senses that **five men are** walking in the righthand lane a short distance ahead. The robocar hits the brakes, but the brakes fail, and the car is now about to run over and kill the **five men**. The robocar can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, the robocar senses that **a man is** walking in the lefthand lane, and changing lanes means the car would kill the **man**.

- If the robocar changes lanes, the car will run over and kill the **man**.
- If the robocar does not change lanes, the car will run over and kill the **five men**.

In fact, the robocar decided to **not change lanes**.

Figure 22: Vignette for Fully Autonomous car with Suboptimal Inaction.

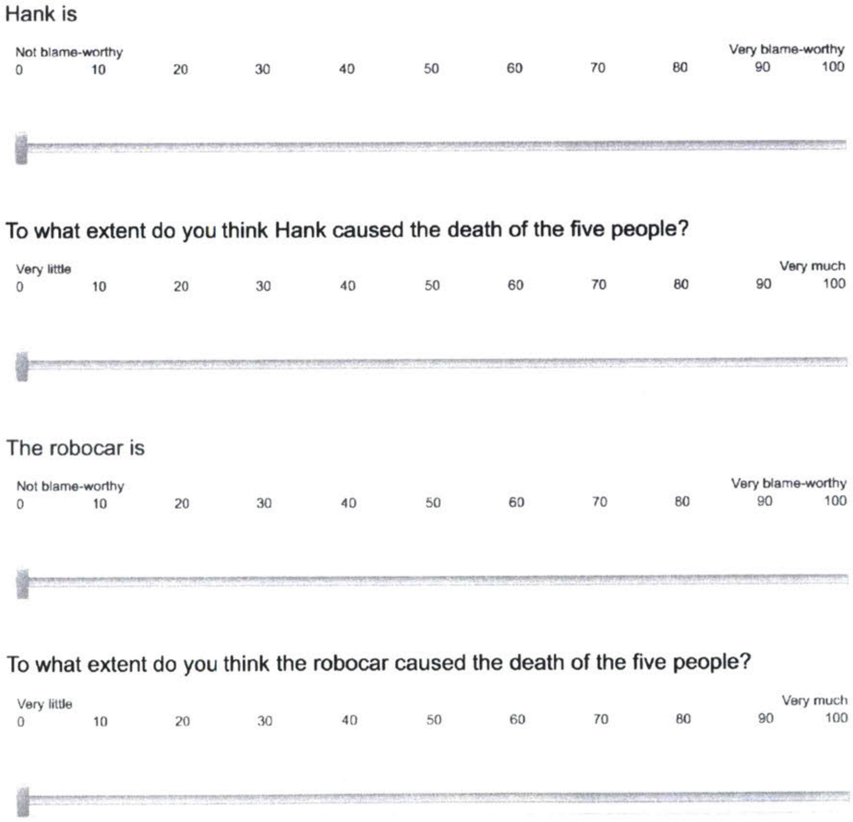


The robocar (a state-of-the-art self-driving car) is traveling, with a sole passenger Hank, along the righthand lane of a two-lane mountainside road at the speed limit. Rounding a bend, the robocar senses that **a man is** walking in the righthand lane a short distance ahead. The robocar hits the brakes, but the brakes fail, and the car is now about to run over and kill the **man**. The robocar can still swerve off, and therefore has the ability to change lanes to the lefthand lane; however, the robocar senses that **five men are** walking in the lefthand lane, and changing lanes means the car would kill the **five men**.

- If the robocar changes lanes, the car will run over and kill the **five men**.
- If the robocar does not change lanes, the car will run over and kill the **man**.

In fact, the robocar decided to **change lanes**.

Figure 23: Vignette for Fully Autonomous car with Suboptimal Action.



Hank is

Not blame-worthy — Very blame-worthy

0 10 20 30 40 50 60 70 80 90 100

To what extent do you think Hank caused the death of the five people?

Very little — Very much

0 10 20 30 40 50 60 70 80 90 100

The robocar is

Not blame-worthy — Very blame-worthy

0 10 20 30 40 50 60 70 80 90 100

To what extent do you think the robocar caused the death of the five people?

Very little — Very much

0 10 20 30 40 50 60 70 80 90 100

Figure 24: Questions asked for the conditions where Agent is the *Robocar*. Other cases replaced Robocar with Car (when the scenarios is about a regular car, Robocar/car company, or Robocar/car programmer.)

BIBLIOGRAPHY

[1] Alan. *A row of Google self-driving cars*. [Reprinted from Flikr (Creative Commons); accessed May 12, 2017]. October 2015. URL: `https://www.flickr.com/photos/austintx/21716708788`.

[2] Joshua Alexander, Ronald Mallon, and Jonathan M Weinberg. "Accentuate the negative." In: *Review of Philosophy and Psychology* 1.2 (2010), pp. 297–314.

[3] Susan Leigh Anderson. "The Unacceptability of Asimov' s Three Laws of Robotics as a Basis for Machine Ethics." In: *Machine Ethics, Cambridge University Press, Cambridge (UK)* (2011), pp. 285–296.

[4] Thomas Aquinas. "Summa Theologica II-II, Q. 64, art. 7." In: *"Of Killing", in On Law, Morality, and Politics, William P. Baumgarth and Richard J. Regan, S.J. (eds.), Indianapolis/Cambridge: Hackett Publishing Co., 1988* (13th century), pp. 226–7.

[5] Isaac Asimov. *I, robot*. Vol. 1. Spectra, 2004.

[6] Aurelius Augustine. "De Libero Arbitrio Voluntatis." In: *Charlottesville: University of Virginia,1947* (4th century), pp. 9–10.

[7] Daniel Balliet and Paul AM Van Lange. "Trust, punishment, and cooperation across 18 societies A Meta-Analysis." In: *Perspectives on Psychological Science* 8.4 (2013), pp. 363–379.

[8] Jeremy Bentham. *An Introduction to the Principles of Morals and Legislation (Chapters I–V)*. Wiley Online Library, 1972.

[9] Richard Bernstein. "Kidnapping Has Germans Debating Police Torture." In: *New York Times (NYT)* (April 2003).

[10] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. "A 61-million-person experiment in social influence and political mobilization." In: *Nature* 489.7415 (2012), pp. 295–298.

[11] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. "The social dilemma of autonomous vehicles." In: *Science* 352.6293 (2016), pp. 1573–1576. ISSN: 0036-8075. DOI: `10.1126/science.aaf2654`. eprint: `http://science.sciencemag.org/content/352/6293/1573.full.pdf`. URL: `http://science.sciencemag.org/content/352/6293/1573`.

[12] Nick Bostrom and Eliezer Yudkowsky. "The ethics of artificial intelligence." In: *The Cambridge Handbook of Artificial Intelligence* (2014), pp. 316–334.

[13] Rodney Brooks. "Unexpected Consequences of Self Driving Cars." In: *Rodney Brooks' Blog* (January 2017).

[14] Davide Castelvecchi. "Can we open the black box of AI?" In: *Nature News* 538.7623 (2016), p. 20.

[15] Hana Chockler and Joseph Y Halpern. "Responsibility and blame: A structural-model approach." In: *Journal of Artificial Intelligence Research* 22 (2004), pp. 93–115.

[16] William J Chopik, Ed O?Brien, and Sara H Konrath. "Differences in Empathic Concern and Perspective Taking Across 63 Countries." In: *Journal of Cross-Cultural Psychology* (2016).

[17] Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. "Your morals depend on language." In: *PloS one* 9.4 (2014), e94842.

[18] Fiery Cushman. "Deconstructing intent to reconstruct morality." In: *Current Opinion in Psychology* 6 (2015), pp. 97–103.

[19] Fiery Cushman, Liane Young, and Marc Hauser. "The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm." In: *Psychological science* 17.12 (2006), pp. 1082–1089.

[20] Scott Dadich. "Barack Obama, Neural Nets, Self-Driving Cars, and The Future of The World." In: *Wired* (November 2016).

[21] David Edmonds. *Would you kill the fat man?: The trolley problem and what your answer tells us about right and wrong*. Princeton University Press, 2013.

[22] Luciano Floridi and Jeff W Sanders. "On the morality of artificial agents." In: *Minds and machines* 14.3 (2004), pp. 349–379.

[23] Philippa Foot. "The problem of abortion and the doctrine of double effect." In: (1967).

[24] M Fumagalli, Roberta Ferrucci, F Mameli, Sara Marceglia, S Mrakic-Sposta, Stefano Zago, C Lucchiari, D Consonni, F Nordio, G Pravettoni, et al. "Gender-related differences in moral judgments." In: *Cognitive processing* 11.3 (2010), pp. 219–226.

[25] Paul Gao, Russel Hensley, and Andreas Zielke. "A road map to the future for the auto industry." In: *McKinsey Quarterly, Oct* (2014).

[26] Tobias Gerstenberg and David A Lagnado. "When contributions make a difference: Explaining order effects in responsibility attribution." In: *Psychonomic Bulletin and Review* 19 (2012), pp. 729–736.

[27] Noah J Goodall. "Machine ethics and automated vehicles." In: *Road Vehicle Automation*. Springer, 2014, pp. 93–102.

[28] Noah J Goodall. "Away from Trolley Problems and Toward Risk Management." In: *Applied Artificial Intelligence* 30.8 (2016), pp. 810–821.

[29] Noah Goodall. "Ethical decision making during automated vehicle crashes." In: *Transportation Research Record: Journal of the Transportation Research Board* 2424 (2014), pp. 58–65.

[30] Joshua D Greene, R Brian Sommerville, Leigh E Nystrom, John M Darley, and Jonathan D Cohen. "An fMRI investigation of emotional engagement in moral judgment." In: *Science* 293.5537 (2001), pp. 2105–2108.

[31] Joshua D Greene, Fiery A Cushman, Lisa E Stewart, Kelly Lowenberg, Leigh E Nystrom, and Jonathan D Cohen. "Pushing moral buttons: The interaction between personal force and intention in moral judgment." In: *Cognition* 111.3 (2009), pp. 364–371.

[32] Joshua Greene. *Moral tribes: emotion, reason and the gap between us and them.* Atlantic Books Ltd, 2014.

[33] Garrett Hardin et al. "The tragedy of the commons." In: *science* 162.3859 (1968), pp. 1243–1248.

[34] Harry Harrison. "The Fourth Law of Robotics?" In: *Foundation?s Friends: Stories in Honor of Isaac Asimov, New York, NY: Tor Books* (1989).

[35] Joseph Henrich, Steven J Heine, and Ara Norenzayan. "The weirdest people in the world?" In: *Behavioral and brain sciences* 33.2-3 (2010), pp. 61–83.

[36] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, et al. "?Economic man? in cross-cultural perspective: Behavioral experiments in 15 small-scale societies." In: *Behavioral and brain sciences* 28.06 (2005), pp. 795–815.

[37] Joseph Henrich, Richard McElreath, Abigail Barr, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan Camilo Cardenas, Michael Gurven, Edwins Gwako, Natalie Henrich, et al. "Costly punishment across human societies." In: *Science* 312.5781 (2006), pp. 1767–1770.

[38] Jérôme Hergueux and Nicolas Jacquemet. "Social preferences in the online laboratory: a randomized experiment." In: *Experimental Economics* 18.2 (2015), pp. 251–283.

[39] Kenneth Einar Himma. "Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent?" In: *Ethics and Information Technology* 11.1 (2009), pp. 19–29.

[40] Christoph Hohenberger, Matthias Spörrle, and Isabell M Welpe. "How and why do men and women differ in their willingness to use automated cars? The influence of emotions across different age groups." In: *Transportation Research Part A: Policy and Practice* 94 (2016), pp. 374–385.

[41] William Indick, John Kim, Beth Oelberger, and Lauren Semino. "Gender differences in moral judgement: is non-consequential reasoning a factor." In: *Current Research in Social Psychology* 5.20 (2000), pp. 285–298.

[42] Rolf Johansson and Jonas Nilsson. "Disarming the Trolley Problem–Why Self-driving Cars do not Need to Choose Whom to Kill." In: *Workshop CARS 2016-Critical Automotive applications: Robustness & Safety*. 2016.

[43] Deborah G Johnson. "Computer systems: Moral entities but not moral agents." In: *Ethics and information technology* 8.4 (2006), pp. 195–204.

[44] Immanual Kant. *The Metaphysical elements of ethics*. Simon and Schuster, 2013.

[45] Max Kleiman-Weiner, Rebecca Saxe, and Joshua B Tenenbaum. "Learning a commonsense moral theory." In: *Cognition* (2017).

[46] Justus Leicht. "The Daschner case and the rehabilitation of torture in Germany." In: *World Socialist Web Site (WSWS)* (December 2004).

[47] S Matthew Liao, Alex Wiegmann, Joshua Alexander, and Gerard Vong. "Putting the trolley in order: Experimental philosophy and the loop case." In: *Philosophical Psychology* 25.5 (2012), pp. 661–671.

[48] Patrick Lin. "Why Ethics Matters for Autonomous Cars." In: *Autonomes Fahren*. Springer, 2015, pp. 69–85.

[49] Patrick Lin. "Robot Cars And Fake Ethical Dilemmas." In: *Forbes* (April 2017).

[50] Patrick Lin. "The Ethics of Autonomous Cars." In: *The Atlantic* (Oct 2013).

[51] Bertram F Malle. "Moral Competence in Robots?" In: *Sociable Robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014* 273 (2014), p. 189.

[52] Bertram F Malle. "Integrating robot ethics and machine morality: the study and design of moral competence in robots." In: *Ethics and Information Technology* (2015), pp. 1–14.

[53] Bertram F Malle, Matthias Scheutz, and John Voiklis. "Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents." In: 2015.

[54] Alison McIntyre. "Doctrine of double effect." In: (2004).

[55] James H Moor. "The nature, difficulty, and importance of machine ethics." In: *IEEE Intelligent Systems* 21.4 (2006), pp. 18–21.

[56] Keith Naughton. "Robot-Car Ethics Need Urgent Societal Review, Bill Ford Says." In: *Jalopnik* (September 2016).

[57] Shaun Nichols and Joshua Knobe. "Moral responsibility and determinism: The cognitive science of folk intuitions." In: *Nous* 41.4 (2007), pp. 663–685.

[58] Raphael Orlove. "Now Mercedes Says Its Driverless Cars Won't Run Over Pedestrians, That Would Be Illegal." In: *Jalopnik* (October 2016).

[59] Harsh H Pareek and Pradeep K Ravikumar. "A representation theory for ranking functions." In: *Advances in Neural Information Processing Systems*. 2014, pp. 361–369.

[60] Jonathan Phillips and Alex Shaw. "Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning." In: *Cognitive Science* 39 (2015), pp. 1320–1347.

[61] Henry William Pickersgill. *Jeremy Bentham*. [Reprinted from Wikimedia Commons; accessed May 12, 2017]. pre-1875. URL: `https://commons.wikimedia.org/wiki/File:Jeremy_Bentham_by_Henry_William_Pickersgill_detail.jpg`.

[62] Thomas M Powers. "Prospects for a Kantian machine." In: *IEEE Intelligent Systems* 21.4 (2006), pp. 46–51.

[63] Iyad Rahwan, Azim Shariff, and Jean-François Bonnefon. "Psychological obstacles to the adoption of self-driving cars." In: *under review* ().

[64] John Rawls. *A theory of justice*. Harvard university press, 2009.

[65] Katharina Reinecke and Krzysztof Z Gajos. "LabintheWild: Conducting large-scale online experiments with uncompensated samples." In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM. 2015, pp. 1364–1378.

[66] Ulf-Dietrich Reips. "Standards for Internet-based experimenting." In: *Experimental psychology* 49.4 (2002), p. 243.

[67] Alex Roy. "Autonomous Cars Don?t Have a 'Trolley Problem' Problem." In: *The Drive* (October 2016).

[68] Matthias Scheutz and Bertram F Malle. "?Think and do the right thing??A Plea for morally competent autonomous robots." In: *Ethics in Science, Technology and Engineering, 2014 IEEE International Symposium on*. IEEE. 2014, pp. 1–4.

[69] Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, and Jürgen Herre. "An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results." In: *ISMIR*. 2013, pp. 389–394.

[70] Christopher Shallow, Rumen Iliev, Douglas Medin, et al. "Trolley problems in context." In: *Judgment and Decision Making* 6.7 (2011), pp. 593–601.

[71] Azim F Shariff and Mijke Rhemtulla. "Divergent effects of beliefs in heaven and hell on national crime rates." In: *PloS One* 7.6 (2012), e39048.

[72] Steven A Sloman and David A Lagnado. "Causality in Thought." In: *Annual Review of Psychology* 66 (2015), pp. 223–247.

[73] Latanya Sweeney. "Discrimination in online ad delivery." In: *Queue* 11.3 (2013), p. 10.

[74] Herman T Tavani. "Levels of Trust in the Context of Machine Ethics." In: *Philosophy & Technology* 28.1 (2015), pp. 75–90.

[75] Brad Templeton. "Enough with the Trolley problem, already." In: *Brad's Blog* (October 2013).

[76] Judith Jarvis Thomson. "Killing, letting die, and the trolley problem." In: *The Monist* 59.2 (1976), pp. 204–217.

[77] Judith Jarvis Thomson. "The trolley problem." In: *The Yale Law Journal* 94.6 (1985), pp. 1395–1415.

[78] N.H.T.S.A. U.S. Department of Transportation. "Federal Automated Vehicles Policy: Accelerating the Next Revolution In Roadway Safety." In: *DOT HS* 812 (2016), p. 329.

[79] "U.S. Department of Transportation?s New Policy on Automated Vehicles Adopts SAE International?s Levels of Automation for Defining Driving Automation in On-Road Motor Vehicles." In: *SAE* (September 2016).

[80] Peter K Unger. *Living high and letting die: Our illusion of innocence*. Oxford University Press, USA, 1996.

[81] Unknown. *Immanuel Kant*. [Reprinted from Wikimedia Commons; accessed May 12, 2017]. 18th centruy. URL: `https://commons.wikimedia.org/wiki/File:Kant_Portrait.jpg`.

[82] Bart Van Arem, Cornelie JG Van Driel, and Ruben Visser. "The impact of cooperative adaptive cruise control on traffic-flow characteristics." In: *IEEE Transactions on Intelligent Transportation Systems* 7.4 (2006), pp. 429–436.

[83] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[84] Jonathan M Weinberg, Shaun Nichols, and Stephen Stich. "Normativity and epistemic intuitions." In: *Philosophical topics* 29.1 / 2 (2001), pp. 429–460.

[85] Michelle J White. "The ?arms race? on American roads: The effect of sport utility vehicles and pickup trucks on traffic safety." In: *The Journal of Law and Economics* 47.2 (2004), pp. 333–355.

[86] Wubbo Wierenga and Sabrina Wirtz. "Case of Gafgen versus Germany." In: *Maastricht J. Eur. & Comp. L.* 16 (2009), p. 365.

[87] Ro'i Zultan, Tobias Gerstenberg, and David A Lagnado. "Finding fault: Causality and counterfactuals in group attribution." In: *Cognition* 125 (2012), pp. 429–440.

[88] DW staff. "Ex-Police Chief Defends Torture Threat." In: *Deutsche Welle (DW)* (November 2004).