INTRODUCTION

Judging Machines

INTRODUCTION

Since Mary Shelley penned *Frankenstein*, science fiction has helped us explore the ethical boundaries of technology.[1] Traumatized by the death of his mother, Victor Frankenstein becomes obsessed with creating artificial life. By grafting body parts, Victor creates a creature that he abhors and abandons. In isolation, Frankenstein's creature begins wandering the world. The friendship of an old blind man brings him hope. But when the old man introduces him to his family and he is once again rejected, he decides that he has had enough. The time has come for the creation to meet his creator. It is during that encounter that Victor learns how the creature feels:

> *Shall each man find a wife for his bosom, and each beast have his mate, and I be alone? I had feelings of affection, and they were requited by detestation and scorn.*

Frankenstein's creation longs for companionship, but he knows that it will be impossible for him to find a partner unless Victor creates one for him. With nothing left to lose, the creature now seeks revenge:

> *Are you to be happy while I grovel in the intensity of my wretchedness? You can blast my other passions, but revenge remains. . . . I may die, but first you, my tyrant and tormentor, shall curse the sun that gazes on your misery. . . . you shall repent of the injuries you inflict.*

Two centuries after Mary Shelley penned *Frankenstein*, we are still unable to graft body parts to create artificial life. But in the world of artificial intelligence (AI), researchers have been creating other forms of artificial "life." One popular format involves the creation of conversational robots, or *chatbots*, who much like Frankenstein's creation, have experienced human scorn.

In 2016, researchers at Microsoft released Tay, an AI chatbot. Just like Frankenstein's creation, Tay was conceived to be beautiful. She was even endowed with the profile picture of an attractive woman. Yet, only sixteen hours after Tay's creation, Microsoft had to shut her down. Tay's interactions with other humans transformed her into a public relations nightmare. In just a few hours, humans turned the cute chatbot into a Nazi Holocaust denier.[2]

As machines become more humanlike, it becomes increasingly important for us to understand how our interactions with them shape both machine and human behavior. Are we doomed to treat technology like Dr. Frankenstein's creation, or can we learn to be better parents than Victor?

Despite much progress in computer science, philosophy, and psychology, we still have plenty to learn about how we judge machines and how our perceptions affect how we treat them or accept them. In fact, we know surprisingly little about how people perceive machines compared to how they judge humans in similar situations. Without these comparisons, it is hard to know if people's judgment of machines is biased and, if so, about the factors affecting those biases.

In this book, we study how people judge machines by presenting dozens of experiments designed to compare people's judgments of humans and machines in scenarios that are otherwise equal. These scenarios were evaluated by nearly 6,000 people in the US, who were randomly assigned to either a treatment or a control condition. In the treatment condition, scenarios were described as concerning the actions of a machine.

In the control condition, the same actions were presented as being performed by a human. By comparing people's reactions to human and machine actions, while keeping all else equal, we can study how who is performing an action affects how the action is judged.

Humans have had a complicated relationship with machines for a long time. For instance, when first introduced, printing was declared demonic by religious scribes in Paris.[3] Soon, it was banned in the Islamic world.[4] A similar story can be told about looms and Luddites.[5] But humans also have a complicated relationship with each other. Our world still suffers from divisions across cultural and demographic lines. Thus, to understand people's reactions to machines, we cannot study them in isolation. We need to put them in context by benchmarking them against people's reactions to equivalent human actions. After all, it is unclear whether we judge humans and machines equally or if we make strong differences based on who or what is performing an action.

In recent years, scholars have begun to study this question. In one paper,[6] scholars from Brown, Harvard, and Tufts explored a twist on the classic trolley problem.[7] This is a moral dilemma in which an out-of-control trolley is destined to kill a group of people unless someone deviates it onto a track with fewer people to kill.* In this particular variation of the trolley problem, the scholars didn't ask subjects to select an action (e.g., would you pull the lever?), but to judge four possible outcomes: a human or a machine pulls the lever to diverge the trolley (or not).

* The exact setup was the following: "In a coal mine, (a repairman or an advanced, state-of-the-art repair robot) is currently inspecting the rail system for trains that shuttle mining workers through the mine. While inspecting a control switch that can direct a train onto one of two different rails, the (repairman/robot) spots four miners in a train that has lost the use of its brakes and steering system. The (repairman/robot) recognizes that if the train continues on its path, it will crash into a massive wall and kill the four miners. If it is switched onto a side rail, it will kill a single miner who is working there while wearing a headset to protect against a noisy power tool. Facing the control switch, the (repairman/robot) needs to decide whether to direct the train toward the single miner or not."

The scholars found that people judged humans and robots differently. Humans were blamed for pulling the lever, while robots were blamed for not pulling it. In this experiment, people liked utilitarian robots and disliked utilitarian humans.[†]

But this is only the tip of the iceberg. In recent decades, we have seen an explosion of research on machine behavior and AI ethics.[8] Some of these studies ask how a machine should behave.[9] Others ask if machines are behaving in a way that is biased or unfair.[10] Here, we ask instead: How do humans judge machines? By comparing people's reactions to a scenario played out by a machine or a human, we create counterfactuals that can help us understand when we are biased in favor of or against machines.

In philosophy, and particularly in ethics, scholars make a strong distinction between normative and positive approaches. A *normative approach* focuses on how the world should be. A *positive approach* describes the world that is. To be perfectly clear, this book is strictly positive. It is about **how humans judge machines**, not about **how humans *should* judge machines**. We focus on positive, or empirical, results because we believe that positive questions can help inform normative work. How can they do this? By focusing our understanding of the world on empirically verifiable effects that we can later explore through normative approaches.

Without this positive understanding, we may end up focusing our normative discussions on a world that is not real or relevant. For instance, empirical work has shown that people exhibit *algorithmic aversion*,[11] a bias where people tend to reject algorithms even when they are more accurate than humans. Algorithmic aversion is also expressed by the fact that people lose trust in algorithms more easily when they make mistakes.[12]

[†] We replicated this experiment using the exact same questions and a sample of 200 users from Amazon Mechanical Turk (MTurk). While we did not find the strong significant effect reported in the original paper, we found a slight (and not significant) effect in the same direction. We were also able to find a stronger effect in a subsequent experiment, in which we added a relationship (family member) between the agent pulling the lever and the person on the track.

Is algorithmic aversion something that we should embrace, or a pitfall that we should avoid?

The social relevance of the question comes into focus only under the light of the empirical work needed to discover it. Normative questions about algorithmic aversion are relevant because algorithmic aversion is empirically verifiable. If algorithmic aversion was not real, discussing its normative implications would be an interesting but less relevant exercise. Because positive work teaches us how the world is, we believe that good empirical work provides a fundamental foundation that helps narrow and focus normative work. It is by reacting to accurate descriptions of the world as is that we can responsibly shape it. This is not because the way that the world is provides a moral guide that we should follow—it doesn't. But it is important for us to focus our limited normative efforts on relevant aspects of reality.

Why should we care about the way in which humans judge machines?

In a world with rampant algorithmic aversion, we risk rejecting technology that could improve social welfare. For instance, a medical diagnosis tool that is not perfectly accurate, but is more accurate than human doctors, may be rejected if machine failures are judged or publicized with a strong negative bias. On the contrary, in a world where we are positively biased in favor of machines, we may adopt technology that has negative social consequences and may fail to recognize those consequences until substantial damage has been done.

In the rest of the book, we will explore how humans judge machines in a variety of situations. We present dozens of scenarios showing that people's judgment of machines, as opposed to humans performing identical actions, varies depending on moral dimensions and context. We present scenarios in which machines and humans are involved in actions that result in physical harm, offensive content, or discrimination. We present scenarios focused on privacy, comparing people's reactions to being observed by machines or by other people. We explore people's preferences regarding labor

displacement caused by changes in technology, outsourcing, offshoring, and migration. We present moral dilemmas involving harm, fairness, loyalty, authority, and purity. We present scenarios in which machines are blasphemous or defame national symbols.

Together, these scenarios provide us with a simple and early compendium of people's reactions to human and machine actions.

In the field of human-robot interactions, people talk about simulated and real-world robot studies.[13] Simulated studies involve descriptions of scenarios with humans and machines like those described in *Frankenstein*. Real-world studies involve the use of actual robots, but they are limited by the range of actions that robots can perform and tend to involve relatively small sample sizes. Simulated studies have the advantage of being quicker and more scalable, which provides a high degree of control over various manipulations. However, because they are based on simulated situations, they may not generalize as well to actual human-robot interactions.

In this book, we focus on simulated studies because they allow us to explore a wider variety of situations with a relatively large sample size (a total of nearly 6,000 subjects, and 150–200 of them per experimental condition). We also chose to do this because these studies resemble more closely one of the main ways in which humans will interact with robots in the coming decades: by hearing stories about them in the news or social media.[14] Still, because our subjects all lived in the US, and because moral judgments vary with time and culture,[15] our results cannot be considered representative of other cultures, geographies, or time periods.

The book is organized in the following way:

Chapter 1 presents basic concepts from moral psychology and moral philosophy, which will help us discuss and interpret the experiments described in the book. It introduces the ideas of moral agency and moral status, which are key concepts in moral philosophy, as well as the five moral dimensions of moral psychology (harm,

fairness, authority, loyalty, and purity). These concepts provide a basic framework for interpreting the outcome of moral dilemmas and studying them statistically. Much of the remainder of the book will focus on exploring how the judgment of an action is connected to a scenario's specific moral dimension and perceived level of intentionality.

Chapter 2 introduces the methodology that we will follow by introducing four sets of scenarios. These involve decision-making in situations of uncertainty, creative industries, autonomous vehicles, and the desecration of national symbols. Here, we find our first patterns. People tend to be unforgiving of AIs in situations involving physical harm, and when AIs take risks and fail. In the self-driving car scenario, we find that people are more forgiving of humans than machines, suggesting a willingness to completely excuse humans—but not machines—when clear accidents are involved. In the creative industry scenarios, we find that AI failures can centralize risks up a chain of command. Finally, we show a scenario involving the improper use of a national flag. This scenario, and another one involving plagiarism, are cases in which people judge humans more harshly, suggesting that people's bias against machines is neither unconditional nor generalized (machines are not always seen as bad). It is a bias that depends on context, such as a scenario's moral dimensions and perceived intentionality.

Chapter 3 focuses on algorithmic bias. The scenarios presented here focused on fairness and involve hiring, admissions, and promotion decisions. They involve a human or machine that either made or corrected a biased decision. We find that people tend to judge humans more strongly in both the positive and negative scenarios, giving more credit to humans when they corrected a bias, but also judging them more harshly when they made a biased decision. We conclude by discussing recent advances in the theory of algorithmic bias, which have demonstrated that simply failing to include demographic information in a data set is a suboptimal way to reduce bias.

Chapter 4 explores issues of privacy by looking at several scenarios involving camera systems used to enforce or monitor public transportation, safety, and school attendance. We also present a few scenarios involving humans or machines using

personal data, including examples along the entire spectrum. In some, we find a negative bias against machines (e.g., school attendance monitoring), while others show no difference between being observed by machines or humans (e.g., camera systems at malls). Yet other scenarios show bias against human observers (e.g., surveillance at an airport terminal), suggesting that the preference for machine or human observers is largely context specific.

Chapter 5 focuses on labor displacement. Here, we compare people's reactions to displacement attributed to changes in technology (e.g., automation), with displacement attributed to humans through outsourcing, offshoring, immigration, or hiring younger workers. We find that in most cases, people react less strongly to technological displacement than to displacement attributed to humans, suggesting that the people in our study tended to be less sensitive to technology-based displacement than to displacement because of other humans.

Chapter 6 brings everything together by using statistical models to summarize the data presented in previous chapters (as well as the additional scenarios presented in the appendix). We find that people tend to be more forgiving of machines in dilemmas that involve high levels of harm and intention and less forgiving when harm and intention are low. In addition, people judge the intention of a scenario differently when actions are attributed to machines or humans. People judge the intention of human actions quite bimodally (assigning either a lot or a little intention to it). Meanwhile, they judge machine actions following a more unimodal distribution—they are more forgiving of humans in accidental scenarios but harsher in scenarios where intention cannot be easily discarded.

In this chapter, we also study the demographic correlates of people's judgment of humans and machines. We find that on average, men are more in favor of replacing humans with machines than are women. People with higher levels of education (e.g., college and graduate school as opposed to only high school) are also a bit more accepting of replacing humans with machines.

Finally, we use data from dozens of scenarios to construct statistical models that help us formalize people's judgments of human and machine actions. The model formalizes a pattern that is prevalent in many scenarios, and, while not 100 percent generalizable, that explains many of our observations: **people judge humans by their intentions and machines by their outcomes.** This finding is a simple empirical principle that explains scenarios like the trolley example presented previously, but many others as well.

Chapter 7 concludes by exploring the implications of the empirical principle presented in chapter 6, and by drawing on examples from academia and fictional literature to discuss the ethical and legal implications of a world where machines are moral actors.

The appendix presents dozens of additional scenarios, which were not part of the main text, but were used in the models presented in chapter 6.

How do humans judge machines? Not the same as humans. We focus more on machines' outcomes, and we are harsher toward them in situations that involve harm or uncertainty, but at the same time, we can be more forgiving of them in scenarios involving fairness, loyalty, and labor displacement. Yet, we still have much to learn. By presenting this collection of experiments, we hope to contribute to a better understanding of human-machine interactions and to inspire future avenues of research.