7

Liable Machines

CHAPTER 7

After lighting a cigarette, Alfred Lanning, declared, "It reads minds all right."[1] Lanning was a recurrent character in Isaac Asimov's science fiction. In this particular story, the director of a plant of U.S. Robots and Mechanical Men was talking about Herbie, a robot with "a positronic brain of supposedly ordinary vintage." Herbie had the ability to "tune in on thought waves," leaving Lanning and his colleagues baffled by his ability to read minds. Herbie was "the most important advance in robotics in decades." But neither Lanning nor his team knew how it happened.

Lanning's team included Peter Bogert, a mathematician and second-in-command to Lanning; Milton Ashe, a young officer at U.S. Robots and Mechanical Men; and Dr. Susan Calvin, a robopsychologist (who happened to be in love with Ashe).

Lanning asked Dr. Calvin to study Herbie first. She sat down with the robot, who had recently finished reading a pile of science books. "It's your fiction that interests me," said Herbie. "Your studies of the interplay of human motives and emotions." As Dr. Calvin listened, she begun to think about Milton Ashe.

"He loves you,"—the robot whispered.

"For a full minute, Dr. Calvin did not speak. She merely stared."

"You are mistaken! You must be. Why should he?"
"But he does. A thing like that cannot be hidden, not from me."

Then he supported his statement with irresistible rationality:

"He looks deeper than the skin and admires intellect in others. Milton Ashe is not the type to marry a head of hair and a pair of eyes."

She was convinced. "Susan Calvin rose to her feet with a vivacity almost girlish."

After Dr. Calvin, it was Bogert's turn. He was a mathematician who saw Herbie as a rival. Once again, Herbie quickly directed the conversation toward Bogert: "Your thoughts . . . concern Dr. Lanning." The mathematician took the bait.

"Lanning is nudging seventy. . . . And he's been director of the plant for almost thirty years. . . . You would know whether he's thinking of resigning?"

Herbie answered exactly what Bogert wanted to hear.

"Since you ask, yes. . . . He has already resigned!"

Bogert asked Herbie about his successor, and the robot confirmed it was him.

But Herbie's story is not that of a robot who bears good news, but that of a mind-reading robot struggling with the "First Law of Robotics." Soon, the scientists and engineers began putting their stories together, discovering that what Herbie had told them wasn't correct. Milton was engaged to be married to someone else, and Lanning had no intention of resigning. Herbie had lied to them, and they wanted to know why.

While the men were pacing around the room, Dr. Calvin had an "aha" moment: "Nothing is wrong with him." Her colleagues paused. "Surely you know the . . . First Law of Robotics?"

Like well-trained schoolchildren, her colleagues recited the first law: "A robot may not injure a human being or, through inaction, allow him to come to harm."

She continued. "You've caught on, have you? This robot reads minds. . . . Do you suppose that if asked a question, it wouldn't give exactly that answer that one wants to hear? Wouldn't any other answer hurt us, and wouldn't Herbie know that?"

Dr. Calvin turned toward Herbie: "You must tell them, but if you do, you hurt, so you mustn't; but if you don't, you hurt, so you must; but. . ."

Failing to deal with the contradiction, Herbie "collapsed into a huddled heap of motionless metal."

* * *

The rise of artificial intelligence (AI) has brought a deluge of proposals on how to regulate it.[2] Tech companies, such as Google,[3] and international organizations, such as the European Commission[4] and the Organisation for Economic Co-operation and Development (OECD),[5] have published plans or convened committees to guide AI regulation.[*] But the global rush to regulate AI is no indication that morality can be reduced to a set of rules.

Almost a century ago, when computation was in its infancy, the mathematician and analytic philosopher Kurt Friedrich Gödel uncovered what is one of the most beautiful

[*] In the case of Google, though, the committee did not last long (S. Levin, "Google Scraps AI Ethics Council after Backlash: 'Back to the Drawing Board,'" The Guardian, 5 April 2019).

axioms of mathematics:[6] the idea that mathematics is incomplete. That incompleteness does not mean that there is a blank space of mathematics that could eventually be filled, but rather that there are truths in a logical system, such as mathematics, that cannot be proved using only the rules within the system. To prove them, you need to expand the system. Doing so answers those truths, but also opens new ones that once again cannot be proved from within. Mathematics is incomplete not because a finite set of proofs is missing, but because every time we try to complete it, we open the door to new and unprovable truths.

Asimov's "three laws of robotics," therefore, may not be a match for Gödel's theorems. And, probably, they did not pretend to be. The story of Herbie is not about the three laws working, but about the first law breaking. This is a common theme in Asimov's writings. Even though he is probably best known for proposing the three laws of robotics, his literature is filled with stories where the laws fail. The story of Herbie is a particularly interesting example involving mundane human desires: a woman liking a man, and a man wanting his boss's job.

There is no reason to believe that a logical system as complex as morality is complete when mathematics is not. In fact, because reducing morality to mathematics may be an impossibility, our moral intuitions may also respond to a logic that is also incomplete. If this is true, trying to reduce machine morality to a set of rules is naive. Before long, either writers like Asimov or robots like Herbie will uncover contradictions. They will find those unproven truths. If morality is incomplete, then it cannot be enforced through obedience.

While scholars have explored a number of moral dilemmas involving machines, some of the most interesting dilemmas are found in recent works of fiction. One of the best examples is the 2018 video game *Detroit*. The game follows the lives of three androids who—after facing a series of moral dilemmas—become human. One of them is Kara, a maid who must care for an abusive dad and his young daughter. She takes care of household chores, serves the father, and also must protect the child. But Kara's

owner pushes these goals into conflict. Kara is expected to obey the abusive dad, but he is the one hurting the daughter. When the contradiction becomes unsustainable, Kara must break one of the rules. It is through this conflict that she becomes a deviant—an android that is no longer obedient to humans, an entity with the free will to choose her own moral path.

Kara chooses to defend the child and is required to fight the dad to do so. The dad throws her around the room violently until Kara manages to shove him into a wall and run away.[†] In doing so, she broke a rule in order to satisfy another, even though most people would agree that in this situation, Kara did the right thing.

But Kara's and Herbie's stories have something in common. They are two examples showing that contradictions can emerge when moral rules are combined with social relationships. Herbie had no problem telling people exactly what they wanted to hear. But when that information was about others, he encountered conflict. Kara could be perfectly obedient to the abusive father and protective of the child. But in the presence of both of them, a moral conflict emerged. For Herbie, the moral trade-off was between lying to avoid immediate harm and causing harm through the future unraveling of his lies (an economist would say that Herbie "infinitely discounted" future harm). For Kara, the contradicting goals were to obey the father and protect the child. Together, both stories illustrate the frustration that moral rules suffer in the presence of social networks. In social groups, Asimov's laws bow to Gödel's theorem.

[†] In the game, there are other possible options—such as shooting the dad—which modify how the subsequent story unfolds.

Videogame *Detroit: Become Human* - Kara Shoots Todd

CHAPTER 7

Responsible Machines

How would you judge Herbie if he were human? How about Kara? What if instead of an android, Kara were a human au pair?

Throughout the last six chapters of this book, we compared people's reactions to a variety of scenarios in which humans or machines were involved. We learned that humans are not generally biased against machines—the direction of the bias (positive or negative) depends on the moral dimension of the scenario, as well as the level of perceived intention and uncertainty. We found that people judge machines more harshly in scenarios involving physical harm, such as the car and tsunami scenarios presented in chapter 2. But we also found situations in which people tend to forgive machines more than humans, albeit slightly. These are scenarios dealing with fairness, like the algorithmic bias scenarios in chapter 3.

When we studied privacy, we found that people are wary of machines watching children, but they are more indifferent to them in commercial settings, such as a mall or hotel. They were also more comfortable with machines in more institutional contexts, such as airport security and citizen scoring.

When we looked at labor displacement, we found that people reacted more negatively to displacement that is attributed to other humans, especially foreign or younger workers. In fact, technological displacement was the option eliciting the least negative reactions.

We then put these various scenarios together in a chapter that described the statistical trends observed across the data. We focused on the harm, intention, and wrongness dimensions of morality and found the moral planes described by these three variables to be different for human and machine actions. Moreover, we found that people judge the intentions of humans and machines differently. People judge humans

following a bimodal distribution, attributing either a lot or a little intention. On the contrary, people judge machine intentions using a unimodal distribution. Machines are not blamed as fully intentional, but they are also not excused as much as humans in accidental situations.

This brings us to what is probably the most poignant observation in our study: **people judge humans by their intentions and machines by their outcomes**. This idea (which is a simplification, of course) is supported by several observations, not only by differences in the judgment of intention. For instance, in natural disasters like the tsunami, fire, or hurricane scenarios, there is evidence that humans are judged more positively when they try to save everyone and fail—a privilege that machines do not enjoy. The idea that we judge machines by outcomes and humans by intentions is also seen clearly in the reduced-form models in chapter 6. These models show that the judgment of machines is, on average, explained mostly by a scenario's level of perceived harm (outcome), whereas the judgment of a human in the same scenario is modulated by the perceived level of intention (and the interaction terms between intention and harm).

Chapter 6 also identified some interesting, albeit mild, correlations between the demographic characteristics of the study's participants and the response functions. People with higher levels of education (college or graduate school compared to high school) were less prone to replace machines with humans and more prone to replace humans with machines, as were men compared to women.

One question that we left relatively unexplored, however, is that of responsibility for machine actions. Our only contribution was the lewd advertising examples of chapter 2, which showed that responsibility shifts toward the most central actors of a chain of command when machines are involved.

Still, the question of responsibility for machine actions is one that has become increasingly important in a world of semi-intelligent machines. It is also an old question that builds on normative frameworks developed to think about product liability.[7]

Product liability law is based on some well-understood principles, such as the ideas of negligence and recklessness. A manufacturer is considered negligent if they *fail to warn of or fail to take proper care to avoid a foreseeable risk.* The requirement to communicate risks is why we find warning labels on products. Failing to take proper is more difficult to characterize, but it usually involves benchmarks with industry standards or common sense. Recklessness is similar to negligence but involves the actor being aware of the risks or avoiding learning about them. Negligence and recklessness can move issues of liability from civil to criminal charges, and yet foreseeing or understanding risk is increasingly complex in a world with machines that are increasingly versatile, complex, and intelligent.

This complexity makes assigning liability more difficult.[8] In principle, liability can be differentially apportioned, but in the case of AI, it may be hard to untangle how much of that liability should be apportioned to data, algorithms, hardware, or programmers. Moreover, AI systems could be quite versatile in their use, could be reprogrammed, or even learn. In general, manufacturers are protected against people using products in wholly unintended ways (such as using an umbrella as a parachute), but in the case of AI, the intended uses could be harder to define; hence, manufacturers may react by restricting the programmability of systems in order to limit their potential liability.[9]

Another idea that should inform the way in which we think about machine responsibility is the idea of *vicarious liability*,[10] which is liability passed to an owner or user (e.g., the liability that a dog owner has for their pet). Some have argued that robots should be treated as domesticated animals[11] because they possess some degree of autonomy but are also not usually ascribed rights or moral responsibilities. Vicarious responsibility could be passed to manufacturers, users, and companies, as we already do in the case of powerful technologies such as cars or explosives. In the case of a car,

manufacturers are responsible for ensuring that they produce safe designs, but drivers are also responsible for the ways they drive and must conform to regulations governing car use and ownership. Still, vicarious responsibility could be passed to an organization. For instance, drivers working for a company transfer a major part of their liability to the company that hires them.

Regardless, the responsibility for machine actions falls to humans. The question is, which humans? The ideas of product liability, vicarious liability, recklessness, and negligence do not provide us with all the answers, but they help us ask the right questions. How much responsibility should be allocated to manufacturers and users? How should responsibility be distributed among hardware, software, and data input? How about mistakes attributed to data generated directly by users, stemming from public sources, or emerging from crowdsourced efforts? How open should these systems be to tinkering and reprogramming? Should AI software be fully open-source, private, or something in between?

Intentions and Outcomes

By looking at hundreds of scenarios, we have learned that people judge humans by their intentions and machines by their outcomes. This simple principle, however, inspires us to think about the way in which humans judge systems more generally, as well as about the role of intention in both human and machine actions.

Beyond machines, people also frequently interact with systems made of people—namely, bureaucracies, like the ones we find in governments or large organizations. Thinking of bureaucracies as machines is not new. In fact, this idea can be traced to the work of Max Weber, the German scholar and philosopher, who is credited for founding the field of sociology in conjunction with Karl Marx and Emile Durkheim. In his treatise on social and economic organization, Weber wrote: "A fully developed bureaucratic

mechanism stands in the same relationship to other forms as does the machine to the non-mechanical production of goods. Precision, speed, clarity, documentary ability, continuity, discretion, unity, rigid subordination, reduction of friction and material and personal expenses are unique to bureaucratic organization."[12]

But while equating bureaucracies to machines may sound metaphorical, the truth is that bureaucracies are designed to be mechanical. Weberian bureaucracies are expected to be impersonal, hierarchical structures governed by rules, regulations, and procedures, and also characterized by a deep division of labor. By all means, they are machines comprised of people who, for the most part, are not empowered to make decisions, but rather are required to act according to an accepted protocol.

Yet, despite being machinelike, many bureaucracies do not appear to be perceived in a similar way as machines. Governments are the epitome of bureaucracies that are judged based on the intentions that people attribute to their leaders. This personification of bureaucratic machines is expressed in the fact that the terms *government approval* and *presidential approval* are sometimes used interchangeably. Despite being machinelike, people often judge government bureaucracies based on the intentions they ascribe to their leaders. The same action, or outcome, can be seen as positive or negative, or as honest or suspicious, depending on whether the person judging the action is politically aligned with the leader.

But the same is not true, or it is true to a lesser extent, for commercial bureaucracies. People's approval of products, like cars, computers, or aircraft, is less influenced by who is the current chief executive officer of the company that makes them. This is probably due to a variety of factors, such as the relative obscurity of business leaders vis-à-vis political leaders and the fact that learning about the quality of a product (e.g., the reliability of a car or computer) is easier than learning about the quality of government services. Nevertheless, the personification of bureaucratic systems has some important implications. First, if we judge bureaucratic systems by focusing too much on the intentions that we assign to their leaders, we can fail to evaluate their

outcomes properly. In this world, inefficient bureaucracies with charismatic leaders often have the electoral upper hand over efficient bureaucracies with uncharismatic leaders. Second, if there were a transition from our current representative democracy to forms of democracy that are either more direct, more digital, or both,[13] we may inadvertently switch our mode of judgment from one focused on intentions to one focused on outcomes. This could be a potentially beneficial change if we can accurately agree on what outcomes are actually desirable and develop accurate ways of measuring them.

Another reflection that is motivated by the principle that people judge humans by intentions, and machines by outcomes, is the role that intention may play on human as opposed to machine learning. Unlike machines, humans are excellent at learning from only a few examples.[14] This ability to generalize correctly may emerge from the ability of humans to transfer knowledge between domains, as well as from our focus on explainable generalizations.[15] Our ability to model the minds of others based on limited observation and to assign intentions to human actions is an example of this ability to learn from only a few examples. Once we have made up our minds about someone and created a mental model of that person's goals and intentions, we can easily interpret any new piece of information in the light of that mental model. This provides us, for better or worse, with a great ability to generalize (i.e., we can draw big conclusions from little information). But this also can limit our subsequent learning because it may be easier for humans to interpret new information in the light of an existing model than to revise the model that we have.

Thus, what makes humans superior learners (our ability to generalize from a few examples guided by mental models built on implied intention) may also make us inferior *unlearners*. Our obsession with intention may be a powerful shortcut for learning, but it also may limit our ability to change our minds once they are made up.

CHAPTER 7

Outro

More than two centuries ago, Mary Shelley penned *Frankenstein*. This groundbreaking work jump-started the genre of science fiction, but it also taught us to think deeply about our relationship with technology. In this book, we have borrowed a page from Shelley's masterpiece by studying people's reactions to dozens of scenarios. We learned that people do not judge humans and machines equally, and that differences in judgment vary based on a scenario's moral dimensions, the characteristics of participants, and a scenario's perceived levels of harm and intention. But we still have much to learn. Our results are moot about a number of important questions, such as: How do people's judgments of machines vary with culture? How do they vary across time? And what are the ethical and legal implications of this new understanding? We leave these and other questions to future research, with the hope that our empirical results contribute to humans' understanding of how we judge machines.