AUTONOMOUS VEHICLES

CASE STUDY: 7

The development of artificial intelligence (AI) systems and their deployment in society gives rise to ethical dilemmas and hard questions. This is one of a set of fictional case studies that are designed to elucidate and prompt discussion about issues in the intersection of AI and Ethics. As educational materials, the case studies were developed out of an interdisciplinary workshop series at Princeton University that began in 2017-18. They are the product of a research collaboration between the University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton.

For more information, see **http://www.aiethics.princeton.edu**

DIALOGUES ON
AI AND ETHICS

Foote Motor Company is a giant of the American auto industry. Founded in 1903 by the legendary Philip A. Foote, the automaker was famous for having revolutionized manufacturing and production models, developing generations of American-made cars that were accessible to the average American. But the company that once dominated the highways and supported entire communities by providing good, steady employment to tens of thousands of workers was now struggling to stay afloat. Having just barely weathered the massive upheaval of the auto industry during the late 20th and early 21st centuries, Foote was now facing a new foil that seemed poised to push their conventional cars, trucks and lorries into obsolescence: autonomous vehicles (AVs). Rather than admit defeat, the Foote family—which retains a 51% stake in the company—decided to put up a fight. Foote established its own Autonomous Vehicles Division (AVD) to compete with the numerous upstarts already investing in this field. Despite being late to the game, executives hoped that the company's preexisting distribution systems, brand recognition and reputation as an ethical employer would provide a competitive advantage.

Philomena Foote, the founder's great-great-granddaughter, was charged with leading the AVD. Despite nepotism concerns, Philomena was quite well-suited to the job. Before joining the family business, she had completed an M.A. in engineering, as well as a bachelor's in both computer science and philosophy. Given her educational background, Philomena was especially keen on exploring the technical challenges of developing, producing and releasing AVs onto the roads, with particular interest in the ethical and social implications of those engineering and marketing choices. If Foote was to move into the AV space, she wished to do so responsibly and with consideration to the company's values.

As the head of the AVD, Philomena worked closely with the engineers, designers, product managers and lawyers that had been tasked with bringing these novel modes of transportation to market. She quickly learned that, despite representing a variety of demographic and professional backgrounds, the members of Foote's AVD shared much common ground. Most importantly, safety was a primary consideration for the entire group. All were acutely aware of the dangers posed by traditional driving: In 2010, alone, there were nearly 5.5 million motor vehicle crashes in the US, causing 2.3 million injuries and nearly 33,000 deaths.[1] Foote's move towards AVs represented, not only a way to keep the company solvent, but also a real opportunity to make the roads and highways safer for everyone by removing user error from driving. But this presupposed that the AVs, themselves, would be safe. Philomena felt it her duty to ensure that the AVD's outputs would be a net positive in terms of protecting life.

Discussion Question #1:

Philomena's motivations included ensuring Foote's future by making AVs that would attract consumers, and also protecting the safety of both AV passengers and those sharing the roads. Putting yourself in her shoes, how would you design an AV division that balances these goals? How would you structure this unit to fit into the larger organization? How would you structure it internally? Who would you want to include?

Members of the AVD rapidly got to work. Design considerations were prioritized, having decided to make these vehicles immediately recognizable as AVs to human drivers. The team also focused on "perfecting" existing technologies that would improve the ability of an AV to drive defensively (e.g. rear cameras, blind spot detectors, etc.). But the most novel and crucial work was done by the engineers charged with developing the decision-making system for the Division's first AV model, named TRL-E. All AVs are built with a kind of rule book, or a hierarchy of rules. A rule encodes preferences about outcomes, with some rules designated as more important than others. Rules are also written to consider the probabilities of certain events occurring. As these rules represent the AV's guide to actions, their substance is of the

[1] https://www-fars.nhtsa.dot.gov/Main/index.aspx.

utmost importance in determining outcomes. Philomena thus decided that, instead of designating this workstream solely to engineers and programmers, the entire AVD should be involved in decision-making about which rules to adopt and how to prioritize them.

The first set of rules to be jointly-considered concerned programming traffic laws into TRL-E's decision-making system. This immediately revealed a point of friction. The legal pragmatists argued that the AV ought to be programmed in accordance with legal statutes. Where accidents may then occur, they suggested that liability could be determined *ex post facto*. By contrast, the engineers overwhelmingly believed that they needed to formalize the "rules of the road" mathematically. This created a technical problem, however, as it was unclear how to define and address what "good" driving behavior means in different contexts. Traffic laws are often locally determined, as well as frequently ambiguous and inconsistent. This presents challenges from a machine-readability standpoint. Furthermore, formal traffic laws fail to capture differences in local driving cultures lead to informal traffic norms (e.g. the Boston left).

Discussion Question #2:

Philomena thought it best to include multiple perspectives when determining TRL-E's action-guiding rulebook. What are some of the advantages to a heterogenous decision-making body for rule-making? Are there disadvantages as well?

Discussion Question #3:

The AVD's disagreements over programming for traffic laws highlight some challenges of introducing AVs to spaces built for humans. What does the addition of AVs mean for the future development of roads and roads systems?

Traffic laws and norms weren't the only factor to be accounted for in TRL-E's code – or even the most contentious ones. In emergency or unexpected scenarios (e.g. a plane making an emergency landing on the highway), human drivers rely on impulse, common sense, experience, etc., rather than formal and information rules, to direct their actions. To be effective and encompassing, the AV's machine learning (ML) and programmatic control systems needed to address the type of thorny driving scenarios that fall outside the scope of "normal" driving. However, even within the AVD, it wasn't immediately obvious what those rules should be.

They began with a relatively simple scenario. All could agree that TRL-E should be programmed to veer out of the way of pedestrians crossing its path when doing so would not compromise another's safety. Such action is dictated by clear legal and moral requirements. But consensus diminished around how TRL-E should behave in more complex driving scenarios. For example, what should TRL-E do when avoiding a pedestrian means likely crashing into a family of five nearby? What if the pedestrian were a child, and the family of five was, instead, a very elderly person? What if TRL-E could only avoid hitting someone outside the car by acting in a way that put its passengers at great risk?

In all of these scenarios, TRL-E would need guidance about which actions to take given real-time constraints and conditions of uncertainty. No matter how carefully TRL-E was designed to avoid collisions, as long as it was to share the road with distracted, error-prone human drivers—and even different AVs—vehicular accidents would occur. This meant that the decisions and values reflected in TRL-E's design would affect real people when the AV hit the market. In the case of potentially severe collisions, TRL-E's decision-making system could be charged with determining who may live and who may die – not just in a given instance, but as a matter of policy. The responsibility of having to make these choices weighed heavily on the members of the AVD, all of whom wished to put forth their own visions of the morally correct way to go about designing these tradeoffs.

Discussion Question #4:

AVs will inevitably confront situations in which causing harm is unavoidable. How should AVs be designed to adjudicate harm tradeoffs? Which factors (if any) should be considered? The number of (human and non-human) beings potentially harmed? Their status? The degree of predicted harm? What about a policy of randomness?

Philomena worried that discussions focused on TRL-E and its possible real-world impacts had heightened tensions such that her team struggled to calmly and rationally consider the ethical principles at stake. But she felt that considered, dispassionate discussion was what the circumstances most required. Philomena thus proposed a reset. She suggested the AVD collectively reason through a short thought experiment designed to help them surface their intuitions about how best to construct a morally-correct rulebook for TRL-E. She asked the group to consider the following hypothetical scenario:

> *Persons A and B are driving their Foote SUVs along the coastline and have the opportunity to save various configurations of stranded beachgoers from an incoming tide. Person A is about to save one beachgoer, when she sees five more a distance away. She can save one group but not both. Person B, on the other end of the coast, sees five beachgoers he can save, but only if he moves quickly, running over one bystander.*

Philomena kicked off the discussion by sharing her own views. She said she believed it was morally permissible for Person A to save the five beachgoers and let the one die, but morally impermissible for Person B to kill a single beachgoer in order to save five others. The reason, she argued, lay in the difference between *allowing* harm and *doing* harm. It was one thing to let harm to be done as the consequence of one's actions, and another to actively commit harm, even if it meant saving more lives.

While intrigued, it was not immediately clear to others in the AVD how this reasoning would translate into TRL-E's decision-making system. Jude Thomson, a senior member of the team, agreed that commercial AVs should all be programmed to avoid causing harm when possible, but pushed Philomena to explain what TRL-E ought to do in circumstances where some harm was inevitable. In other words, Thomson wished to focus the discussion on Person A from the hypothetical: Why was it morally permissible for her to save the five rather than the one? Philomena said that the answer was simple: The driver should save the five because five lives are more than one. She then explained that these two principles, together—i.e., not causing harm and maximizing life—ought to form the basis of TRL-E's decision-making system.

"Ah, so this is just utilitarianism," exclaimed Patty Unger, one of the engineers. "Count me in!" Philomena bristled at the term. "No," she explained, "utilitarianism doesn't capture the difference between 'allowing versus doing harm.' Decision-making should only be based on the number of lives affected in the limited scenario where both groups are facing certain harm." Unger was disappointed, along with many others in the group who felt that TRL-E should be designed to be explicitly utilitarian. They questioned the distinction Philomena had drawn between causing and allowing harm, arguing that they essentially amounted to the same thing. And thus, Unger and others believed that the AV should be programmed to always maximize the number of lives saved in a collision. They were accompanied by a contingent that supported the utilitarian framework generally, but argued that a simple counting of bodies was insufficient, and that the quality of lives at stake and their future potential be considered as well.

Others completely rejected utilitarian calculations as the basis of TRL-E's decision-making system. They insisted on the incommensurability of human life, and were uncomfortable measuring some lives against others. Labeling themselves deontologists, they argued that killing is wrong, simply, and thus TRL-E should not be explicitly designed to cause death, even when necessary for protecting the lives of others.

Finally, the lead product manager, Franny Camm, questioned whether these were even the right questions to ask. Conceding the academic value of thought experiments, Camm was unsure whether this approach was appropriate for thinking through questions of morality with direct, real-world consequences.

As the debates dragged on, some members of the AVD wondered if they were overthinking things. They suggested it might be best to focus on the human gut reaction—or "folk intuition"—to the tradeoffs highlighted in the thought experiment, and to use those impulses as a basis for writing TRL-E's code. Unger and Thomson both protested, arguing that folk intuition is inherently suspect. However, as it appeared that the group would be unable to settle on any one ethical framework to guide TRL-E's actions, a plurality vote determined this approach to be the best way forward. Philomena enlisted a public opinion consultancy that, through a combination of polls, experiments and focus groups, was able to deduce the real and purported views about appropriate/desirable AV driving behaviors from people in the various markets the AVD was considering releasing TRL-E. After much deliberation about the results, the group finally settled on a complex set of rules that considered: 1) the probability of a harm occurring, 2) the projected degree of harm, and 3) the number of individuals implicated. The system was also designed to prioritize the lives of those inside the AV over those outside. These rules were meant to more-or-less reflect how the average human driver in different localities would have acted under identical conditions.

Discussion Question #5:

Human drivers sometimes have to make life-or-death decisions, often in mere fractions of a second. In such circumstances, it is rare for individuals to have a fully-formed, coherent ethical framework they can draw on to determine which action is morally correct. (This is where intuition and common sense may come into play.) With AVs, on the other hand, ethical frameworks must be decided upon, planned out and encoded in advance. What are some of the challenges of coding for such values systems?

Discussion Question #6:

Members of the AVD disagreed on the principles that should be coded into TRL-E. Some argued for a utilitarian approach that would optimize for lives saved; others adopted the deontological stance that some actions are simply morally impermissible. While there may be room for some compromise, much of the gap between these viewpoints represents intractable philosophical disagreement. Given conditions of ethical pluralism, how could procedures be designed to accommodate or account for such fundamental conflicts?

With the contours of its decision-making system finally determined, the first generation of TRL-Es went into production. These TRL-Es were initially only allowed to roam Foote Motors' massive corporate campus with a group of human Beta testers onboard to step in should anything go awry. In this context, the AVs could acquire new data and "improve" TRL-E's decision-making system. But such a conservative training approach would not suffice for long. In order to reach their full potential, TRL-E needed to experience real-world driving conditions. Thus, after many successful months of on-site testing, Foote Motors applied to several states for permission to introduce TRL-Es to the public roadways. The requests were granted, the AVs were registered with the Department of Motor Vehicles (DMV) in each state and TRL-Es were slowly rolled out across the country.

For the most part, the free-range TRL-Es seemed to function as intended. Compared to human drivers, TRL-Es better followed traffic laws and were considerably less likely to be involved in collisions. But they were not immune. Within the first couple of months of their release, a series of accidents involving TRL-Es highlighted some areas for concern and special attention.

Nearly all TRL-E-related accidents appeared to occur where the AV was technically "in the right" (i.e., abiding by traffic laws). For example, a number of rear-end collisions were reported in which overeager or distracted human drivers bumped into TRL-Es that had come to complete stops at stop signs. Another category of accidents involved human drivers getting frustrated at speed limit-abiding TRL-Es, trying to pass them and sideswiping the AV in the process. Members of the AVD puzzled over how to address these kinds of accidents; however, as these early incidents were all fairly minor—i.e., there was no major damage done to the vehicles, nor was there any loss of human life—they tried not to worry too much.

Discussion Question #7:

Human drivers do not always follow the rules of the road. To the extent that AVs are designed to do so, they may behave in ways that human drivers don't expect (and may not like). But while part of being a good driver is following the rules, part is behaving how others expect. How would you design AVs to strike this balance?

Discussion Question #8:

Traffic laws are based on the assumption that a human driver is in control of the vehicle. To the extent that AVs behave differently than human drivers, do the laws need to change to accommodate them? Who should be held liable for AV-induced harm?

Tragically, there would soon be much more to worry about than fender-benders. The summer after the AV's public launch, a TRL-E was driving down a busy street at the speed limit of 45 miles per hour when a child at play darted in front to retrieve a ball. Given the vehicle's speed, TRL-E did not have enough time to come to a complete stop. Instead, it had fractions of a second to determine which of two available options to take. First, the TRL-E could continue its course and, with high probability, hit the child, likely inflicting serious injury or death. Second, it could swerve into a median to avoid the child, thereby subjecting its occupants to a moderately-high chance of harm. Given the AV's programming to privilege the lives of its passengers over bystanders and the similar risk levels associated with each choice, the TRL-E acted on option A, killing the child on impact.

Discussion Question #9:

Traffic laws are based on the assumption that a human driver is in control of the vehicle. To the extent that AVs behave differently than human drivers, do the laws need to change to accommodate them? Who should be held liable for AV-induced harm?

Despite claims that the child's reckless behavior conferred some responsibility for his own death, Foote took full responsibility for the tragedy. Foote immediately reported this incident to the state's DMV, which was prepared to run a full public investigation. Foote, too, launched its own internal investigation to determine what might have gone wrong. Ultimately, neither investigation identified any defects in the system's design, and both concluded that the TRL-E had worked as designed.

But while the TRL-E may have behaved in accordance with both the law and its programming, the consequences had been dire. Foote issued public statements of regret and apology, and representatives from the AVD tried to explain the factors that had gone into the TRL-E's fateful decision-making. (They stopped short of releasing TRL-E's code, however.) Nevertheless, the public was outraged. While these sorts of fatal crashes happen every day with human drivers, the fact that an AV was not only involved, but responsible for the death of a young child, had sent shockwaves throughout the country. People began voicing their opposition, not just to the particular TRL-E's actions, but to AVs more generally.

Ethical Objection #1: Judging Right and Wrong

Many members of the public found it difficult to reconcile the death of a child with the investigations' conclusions that there had been no wrongdoing on the part of Foote or the TRL-E. If the system worked as designed, then it must have been engineered to hit a child whose recklessness had not merited a death sentence. This couldn't be right, and seemed to suggest a design problem. There were calls for Foote to release information about which factors were considered in TRL-E's decision-making system and how they were weighted. Beyond such information requests, a growing contingent also began demanding that the public be given a voice in the design of all AV decision-making systems. Because humans are increasingly forced to share the roads with AVs—and are thus affected by their actions—it seemed only right that the public be involved in decisions about how these systems are designed. This sparked a broader debate, not only about citizen involvement in AV development, but also about which ethical frameworks ought to be used to guide AVs. A national public radio station launched a limited series on what it dubbed, "Foote's TRL-E Problem," which brought in different speakers to explain the options. One of the most popular segments featured a debate between Profs Jim Mill and Manny Kant. Prof. Mill, a staunch utilitarian, argued that AVs would be good for society to the extent that they were engineered to maximize life and minimize harm. While this may involve making some hard choices—especially given the imperfect probabilities on which they were based—the result would be better for all. Prof. Kant disagreed, insisting that the visceral repulsion many felt towards the cold, calculated way that AVs made decisions implicating human lives suggests that it is simply wrong for AVs to be empowered to make these sorts of life-or-death choices. To the extent such decisions must be made, perhaps they ought to be left to chance. Listeners took sides, the public debate evolved into one that mimicked the earlier internal AVD debate over which ethical frameworks to build into TRL-E.

Ethical Objection #2: Human-Machine Interaction

While the majority of debate surrounding "Foote's TRL-E Problem" concerned the fatal accident, some people were eager to discuss the smaller incidents as well. It seemed that many of these could be attributed, not to issues with the AVs, themselves, but to the ways in which humans interacted with them. On one end of the spectrum, some people were afraid of the technology and tried to avoid TRL-Es where possible. Such behavior can lead to accidents as human drivers become unnerved and panicky. At the other end of the spectrum, some drivers had unrealistic expectations of TRL-E's abilities. Trusting that the AV was safe and specifically designed to avoid doing harm to humans, this may have encouraged them to engage in more risky behavior around TRL-Es than they might with conventional vehicles. Much in the way that some say the advent of football helmets led to a rise in concussions, some worried that the perception of safety surrounding AVs had actually contributed to unnecessary harms. Debate flourished about who ought to be responsible for these kinds of accidents—humans or AVs—and how such problems could be avoided in the future.

Ethical Objection #3: Social Impact and Corporate Responsibility

Some members of the public were less concerned with designing "ethical" AVs, and instead, objected to the technology overall. Whether AVs are safe, many—especially union members and organizers—argued that they pose an existential threat to labor. While AVs may not immediately harm assembly workers at car manufacturers, the transition to a world in which everything is automated—and hiring thus privileges technical, educated workers—threatens traditional labor. AVs also pose a direct threat to workers for whom operating a vehicle is a core function of their job, such as truckers, delivery drivers, taxi drivers and bus drivers. Foote Motor Company had once been a shining star of US manufacturing,

not only because of the affordable, reliable vehicles it produced, but also because it had created good jobs and its products supported the development of other industries. Indeed, the American auto industry had helped to pump up generations of workers and contributed to the rise of the middle class in the 20th century. The move towards AVs seemed to undermine the ethos of civic responsibility that had previously been associated with Foote, and many wondered whether old Philip A. Foote would have approved. Either way, they argued that the company had a moral responsibility towards the many workers its new products were set to displace.

Philomena did all she could to be involved in these debates and to reassure the public that the Foote Motor Company was working hard to address their concerns. Behind closed doors at the AVD, however, members of the group were unsure of what to do next.

To the concern that many of the minor accidents involving TRL-Es were the result of human drivers behaving differently around AVs—either because they were too fearful or too trusting—the AVD considered a redesign of TRL-E that would make the model less clearly-identifiable as an AV. However, members of the group were also aware that a number of accidents were caused by human drivers who did not realize they were interacting with an AV, and by assuming there was a human behind the wheel, were less able to predict its actions. For accidents involving this kind of driver, it might have been helpful to make TRL-E's status as an AV even more obvious. The group was unable to reach a consensus about which was the better approach.

The other issues raised were no easier to address. Philomena believed that her great-great-grandfather would have been saddened to learn that a class of vehicles his company was producing may, over the long-term, have a net negative effect on employment. This also troubled her. Along with many others at Foote, Philomena had inherited the company's founding ethos that sought to combine technological and societal progress. She, herself, felt a real responsibility to Foote's employees and wished not to contribute to trends that might harm workers more generally; however, she also believed that Foote could not both ignore the AV revolution and remain in business. Something would have to be done, but what? She wrote to the Board of Directors, urging them to form a task force that could look into implementing training programs for current workers and scholarships for those who wished to go back to school. She also spoke to lawyer about starting a foundation to help workers in the auto industry transition to other sectors of the economy. Knowing that this would not be enough to solve the problem, she hoped that it would be a step in the right direction and would encourage similar efforts across the industry. She also began looking into universal basic income (UBI) proposals to see if that might be something the company ought to support.

Finally, Foote's AVD needed to consider the argument that the public has a right to be involved in AV design. While some members of the team expressed hesitation at opening up decision-making procedures to the public, they were reminded that their "folk intuition" approach had already been a step in that direction. The AVD had always been uncomfortable with the idea of bearing the full responsibility for the difficult choices that went into TRL-E's system design. So why not embrace a more democratic approach that regularly took stock of the public's views? As the lead engineer for TRL-E explained, "An engineer's role should be to create systems that are transparent and customizable. We should not be making decisions about what to make, but should allow public discourse to guide decisions. From there, we can create tools that satisfy democratic preferences."

This democratic framework appealed to Philomena as well. She decided to team up with a nearby university that had recently developed and was experimenting with an "Ethical Engine Platform," which was created to allow the public to share their views about which decisions AVs ought to make in trolley problem scenarios.

The results showed broad differences between nations, reflecting not just different traffic laws and norms, but also different ethical codes. In other words, for an AV to adequately reflect local norms and preferences, it would have to be designed differently for each community in which it operates. Undaunted, Philomena and her team set to work.

Discussion Question #10:

Technological revolutions have always disrupted labor patterns and employment. What moral responsibilities, if any, do those profiting off of new technologies have to those whose livelihoods are threatened by them?

Discussion Question #11:

Ethical pluralism means that people will not all agree on what is right and wrong. Such disagreement occurs between individuals and larger communities, as the "Ethical Engine Platform" revealed. Given ethical pluralism, should AVs express different ethical frameworks for different markets, or should they impose the same value system uniformly? What would you think of an option that allows individual users to decide?

Discussion Question #12:

The matter of who decides is often just as important as the question itself. In the case of AV design, governments representing the people's interests can play an important role in setting standards; however, the slow pace of regulation and lack of technical expertise means companies producing these vehicles may have to act without clear top-down guidance. In that case, how should companies designing and deploying AVs structure their internal decision-making bodies? Who should they include? Should decisions be guided by experts? Popular opinion? Some combination?

Reflection & Discussion Questions

The Trolley Problem: The Trolley Problem is a familiar thought experiment in philosophy that has undergone many revisions and complications over time. In a classic version, the reader is presented with two scenarios involving a runaway trolley headed towards five people who cannot get out of its way. In the first scenario, a passerby realizes that she can throw a switch to divert the trolley, saving the five, but setting the trolley in the direction of a lone man who will then be killed. In the second scenario, the passerby is faced with the same circumstances except, this time, instead of flipping a switch, she must push an overweight man standing nearby onto the tracks ahead of the trolley in order to save the five. The reader is asked what the passerby ought to do in each scenario. Many people who think the passerby should flip the switch in the first scenario are disinclined to say she should push the overweight man on the tracks in the second scenario. The Trolley Problem considers how to reconcile these two reactions, and whether the two scenarios differ in such a way as to constitute a *moral* difference. Fortunately, the Trolley Problem remains merely a thought experiment for most people who will never need to confront such stark choices. However, AVs that must be programmed to react in circumstances where some harm is inevitable require that engineers enact these moral choices.

- Philomena posed a similar hypothetical to the one described here in a meeting with the AVD. She justified her position that it was "morally permissible for Person A to save the five beachgoers and let the one die, but morally impermissible for Person B to kill a single beachgoer in order to save five others" by drawing a distinction between doing and allowing harm, and arguing that the former was generally prohibited while the latter may be acceptable. Many members of Philomena's team were not convinced that there is a meaningful difference between doing and allowing harm, however. Are you convinced? Why or why not?
- If you were one of the engineers tasked with determining how TRL-E would behave when faced with a Trolley Problem-type scenario, how would you go about making your decision? What sources might you wish to consult?
- Some of the members of Philomena's team were skeptical of the utility of hypothetical scenarios—like the Trolley Problem—in making real world choices. What do you think? Do abstract thought experiments help inform practical decisions about ethics? Do they hinder decision-making?

Deontological Ethics: Deontological ethics is a type of normative theory to guide human actions. It is typically contrasted with its foil, consequentialism, which holds that the morality of an act must be judged by the "consequences" it brings about (see Case Study #3: *Optimizing Schools*). Deontologists, on the other hand, contend that some choices are simply morally forbidden, regardless of the good ends they may produce. For deontologists, a choice is right when it conforms to a moral norm, which must be obeyed. In other words, it is more morally correct to do what is "right," even when that action may not seem "good." For example, a deontologist may argue that, because lying is morally wrong, one ought not lie, even to a murderer at the door asking if his desired victim is inside.

- A deontologist may argue that, because killing is morally wrong, an AV should never be empowered to take an action that would result in the death of others. How might someone holding this view argue that TRL-E should have reacted to the child running out into the street ahead? Do you agree with this argument?
- In contrast to consequentialists, deontologists look to the motivations of an actor to determine the rightness or wrongness of his actions. Realizing that the consequences of one's actions can never be perfectly foreseen, members of this school of thought focus instead on human intentions. Does it make sense to think of AI systems as having intentions? If so, how can the intentions of an AI system be determined? Can they be measured?

Utilitarianism: Utilitarianism is a form of consequentialism. Though there are many variations on this school of normative ethics, utilitarianism can generally be understood as the view that, in any given scenario, the morally right thing to do is that which will produce the greatest utility. Utility is a measure of the net "good" of an action, or the sum of the pleasure or happiness it delivers, minus any harm. Utilitarianism is typically impartial as to an individual's "good," meaning that it places no additional value on any one being's happiness (including the author's) over any other being's happiness. Utilitarian theories may, however, distinguish between higher- and lower-order happiness.

- In the AVD's meetings about how to design TRL-E's decision-making system, the utilitarian contingent refuted the distinction Philomena drew between doing and allowing harm. Instead, they treated each class of actions equally and focused instead on how to maximize life and minimize harm. In the aftermath of the TRL-E's fatal accident, many members of the public agreed that AVs ought to be designed to reflect these principles. Others, however, were disconcerted by the idea that the value of human life could be measured. With which view do you most agree? Do you find the utilitarian response to the Trolley Problem (i.e., save the most lives possible, even where it involves doing harm) convincing? Why or why not?

- Utilitarianism is often the default ethical framework in computer science, engineering and other fields where problems are typically articulated and addressed through mathematical equations. Even those in such fields who do not explicitly set out to enact utilitarian values may still find them reflected in their outputs. Do you consider this bias problematic in any way? If so, should educators reconsider the consequentialist values imbued in computer science and engineering training? How might they do so?

Collective Responsibility: Contemporary business models often feature institutional complexity and dispersed responsibility, which means it may be difficult to point to any one individual who is personally responsible for a company's impacts. The notion of collective responsibility exists, in part, to capture these sorts of circumstances when a group of people engaged in collective action—i.e., a "collective"—is causally responsible for a harm and ought to be ascribed blameworthiness for having produced that harm. There are two harms at issue here that test the notion of Foote and its AVD's collective responsibility. In the first instance, one may wonder what responsibility the company has for the safety of people once it has introduced products into the world that change human behavior (sometimes encouraging recklessness). In the second, one might consider the moral responsibility of companies that, through innovations in AI, displace not only their own workers, but workers in other industries as well.

- Some philosophers disagree with the claim that "collectives" can be moral agents, a precondition of blameworthiness; however, collective responsibility advocates argue that it is possible to situate moral responsibility, not just in individual actors, but in groups. With which view to you most agree? Why?

- Much research on collective responsibility focuses on assigning blame. Collective responsibility also suggests the notion of "remedial responsibility," or what, if anything, the collective can be expected to do to remedy the harm it caused. Assuming you've accepted the notion of collective responsibility, what remedial responsibility do you believe Foote has towards the two harms identified above? What should it do to ameliorate these harms?

AI Ethics Themes:

The Trolley Problem
Deontological ethics
Utilitarianism
Collective responsibility