Chapter 16

A Kantian Course Correction for Machine Ethics

Ava Thomas Wright

The central challenge of "machine ethics" is to build autonomous machine agents that act morally rightly.[1] But how can we build autonomous machine agents that act morally rightly, given reasonable disputes over what is right and wrong in particular cases? In this chapter, I argue that Immanuel Kant's political philosophy can provide an important part of the answer. The problem that Kant's political philosophy attempts to solve is how to rightfully resolve reasonable disputes between moral equals in cases where their rights and duties with respect to each other come into conflict. Kant argues that only a legitimate *public* authority under laws to which everyone can consent can settle such disputes in a way that respects everyone's moral equality. The judgments of the legislative, executive, and judicial institutions of such an authority, therefore, take moral *priority* over private ethical opinions about how to resolve such disputed cases. Hence to act morally rightly, autonomous machine agents must, first of all, act in accordance with justice and legitimate public laws.

The chapter has four main sections. In the first section, I criticize what I regard as a misguided approach to the problem of reasonable disputes in machine ethics, which is to build agents that act in accordance with what most people would prefer the agents to do in controversial cases. This approach would result in *immoral* machines that fail to respect the moral equality of persons. In the second section, I set out Kant's approach to the problem of reasonable disputes. I review Kant's statement of the problem, his solution, and its main implication, the Kantian priority of right.

In the third section, I show how appeal to the Kantian priority of right resolves the conflicts between rights in the famous "trolley problem," which has attracted significant attention in machine ethics because self-driving cars may face analogous conflict cases. Finally, in the fourth section, I consider how autonomous machine agents should handle unresolved conflicts between narrow legal obligations, since appeal to the priority of right cannot resolve them. I conclude with a summary of my main claims.

Introduction: (Im)moral Machines

Machine ethics traditionally has been approached from one or some combination of three main moral theoretical frameworks: (1) Consequentialism (e.g. utilitarianism), (2) virtue ethics (e.g. Aristotle's virtue ethic), or (3) deontology (e.g. Kantian ethics).[2] Each moral theory may indicate a different action to take in particular situations, and applied ethicists are often tasked to work out how each theory would resolve controversial cases. But dispute runs wide and deep in ethics. While some agreement on the morally right action in particular cases might sometimes be achieved among those who accept the same moral theoretical framework, disputes over which framework to adopt in the first place are notoriously intractable. Designers of autonomous machine agents thus seem left with the

DOI: 10.4324/9781003189466-20

perplexing problem of which moral theory to adopt as well as how to implement behavior that conforms with that theory in the agent.

One answer is to build agents that act in accordance with what most people would prefer the agents to do in controversial cases. In the online "Moral Machine Experiment," millions of subjects were asked what a self-driving car should do in various accident scenarios where its only choices were to swerve or maintain its lane (Awad et al. 2018). Subjects were asked to decide who lives or dies among characters who varied by nine attributes such as their age, gender, whether they were jaywalking, whether they were passengers or not, etc. Subjects' decisions in these scenarios were then aggregated and analyzed in order to determine the relative strength of collective ethical preferences with respect to these attributes, all other things being equal. While the strongest such ethical preferences found were to spare more rather than fewer lives, and to spare humans over animals, the Moral Machine Experiment also found strong ceteris paribus preferences to spare those of higher status over those of lower status, younger over older people, females over males, and the "fit" over the "large" (Awad et al. 2018, 61–62).

When reporting these results, the authors of the Moral Machine Experiment did not argue that autonomous vehicles should be programmed to act in accordance with the popular ethical preferences they had collected. Their aim, instead, was to initiate a "conversation" that might help us decide as a society what self-driving cars should do in such controversial cases (Awad et al., 63). In a companion paper, however, some of the original authors of the Moral Machine Experiment review the philosophical literature on moral conflicts and raise the question, "[H]ow can society agree on the ground truth [correct ethical decisions]—or an approximation thereof—when even ethicists cannot?" (Noothigatthu et al. 2018, 1). They then propose a solution: "We submit that [moral] decision making can, in fact, be automated, even in the absence of… ground-truth principles, by aggregating people's opinions on ethical dilemmas" (Noothigatthu et al. 2018 1).[3]

This proposal is both naive and misguided. It is naive of a long social contract tradition in political philosophy that addresses the problem of reasonable disputes in such cases; and it is misguided because it ignores two obvious objections.[4] First, popular ethical preferences about how to resolve controversial cases may be wrong. There is no necessary connection between what is ethical and what a popular majority in society believes is ethical; for example, the majority's preferences in Nazi Germany or in the antebellum American South were egregiously unethical. Nor are the preferences of a global majority such as those that the Moral Machine Experiment or Delphi attempts to capture necessarily ethical; in past epochs, a global majority likely would have rejected the equal rights of ethnic, racial, and religious minorities, women, and LGBT people, as well as many other modern values; indeed, a global majority may not accept such rights and values today.[5]

Second, even if a popular majority were correct about how controversial conflict cases affecting rights should be resolved, the direct translation of the majority's raw ethical preferences into action in such cases would often be unjust. The rationale for doing so seems rooted in a vague background sense of the moral legitimacy of democratic rule. But the legitimacy of majority rule in a democracy depends, at a minimum, on its respect for rights of freedom and equality, as well as a number of other institutional and procedural safeguards to establish the rule of law such as representative government, the separation of powers, due process guarantees, etc. The tyranny of a majority acting outside the rule of law is no more morally legitimate than the tyranny of a king. Consider again the preferences found in the Moral Machine Experiment: While acting on popular preferences to spare more rather than fewer lives, or humans over pets, seems morally unobjectionable, acting on preferences to spare higher over lower status people, the fit over the large, females over males, or the young over the old are morally problematic. These latter preferences raise a strong intuition that

acting on them would fail to respect the moral *equality* of persons. Autonomous vehicles that acted in accordance with them in accident scenarios would, therefore, act unjustly.

Machine ethics seems to me to need a course correction. The direct application of popular ethical theories or opinions to determine how autonomous machine agents should act in cases of conflict subject to reasonable dispute is illegitimate and ill-advised. Machine ethics should turn, instead, to meet the moral demands that justice and the rule of law impose on us to build autonomous machine agents that act in ways that respect the freedom and moral equality of everyone.

The Kantian Priority of Right

The authors of the Moral Machine Experiment raise the right question for machine ethics, even if their proposed answer is misguided. Recall their question, "[H]ow can society agree on the ground truth [correct ethical decisions]—or an approximation thereof—when even ethicists cannot?" The cases of concern are those in which our rights with respect to each other are in *conflict*, and where the correct ethical resolution of that conflict is subject to *reasonable* dispute. If there were no such conflict cases, or if there were no reasonable disputes over how to resolve them, then we could just consult ethicists to clarify their correct resolution. There would also be no problem to solve if everyone were not morally *equal*. What moral equality means is that no one has any more natural moral authority than anyone else does to rule over others. If someone had the natural moral authority to settle disputes over our rights—for example, a divinely ordained king—then we would have a duty to defer to the judgment of that superior authority, even if we were to reasonably disagree with it.

This problem of *reasonable disputes* between *moral equals* over their *respective rights and duties in conflict* cases is precisely the problem of justice that Kant's political philosophy attempts to solve. How can we resolve such disputes rightfully, in a way that respects everyone's moral equality? In the next three subsections, I set out Kant's statement of the problem, his solution, and its main implication, the Kantian priority of right.

The Problem of Justice: Reasonable Disputes over Natural Rights

Kant defines the "innate right of freedom" as follows:

> *Freedom* (independence from being constrained by another's choice), insofar as it can coexist with the freedom of every other in accordance with a universal law, is the only original right belonging to every [person] by virtue of [her] humanity.
>
> (DR: 6:237)

Freedom is thus "independence from being constrained by another's choice," and the *right* of freedom is that freedom systematically limited by everyone else's equal freedom under a universal law. Kant elaborates that the innate right of freedom includes a number of constituents such as "innate *equality*" that are "not really distinct from it" (DR: 6:237–8). The innate right of freedom, moreover, entails that we have powers to acquire rights in property, contract, and in status relations such as marriage or parenthood, Kant argues (see DR: 6:250–1).

The problem that Kant's theory of justice confronts is how to rightfully resolve disputes over the order, scope, and limits of our rights in a system of equal freedom under a universal law. The problem arises as a result of two main claims: First, according to Kant, reason cannot by itself completely specify what our rights and duties with respect to each other

are (DR: 6:312). While reason may be capable of determining our respective rights in clear cases, the precise order, shape, and scope of our rights with respect to each other in many or most cases cannot be determined by appeal to reason by itself.[6] Reasonable disagreements over our rights are thus unavoidable. The problem is acute in cases of dispute over acquired rights to property or in contract. Kant stresses that any such rights are merely "provisional" in a state of nature lacking a public authority because they are subject to reasonable dispute (DR: 6:264).

Second, the innate right of freedom is equivalent to innate *equality*. What innate equality means, Kant says, is that everyone has her "own [natural] right to do what seems right and good to [her] and not to be dependent on another's opinion about this" in cases of reasonable dispute over rights (DR: 6:312). Like others in the social contract tradition, Kant rejects the superior natural right of divinely ordained kings to rule over others. No one individual or group has the natural moral authority to unilaterally define everyone's rights and duties with respect to others (i.e., legislate them), or to enforce them (i.e., execute them), or to resolve disputes over them (i.e., adjudicate them).

Kant concludes that morally rightful relations with others are impossible in a state of nature. While the state of nature may not devolve into a Hobbesian civil war, "it would be a state devoid of justice (status justitia vacuus)," Kant says, because "when rights are in dispute (ius controversum), there would be no judge competent to render a verdict having rightful force" (DR: 6:312). Hence, rightful relations in the state of nature are impossible, Kant says, even if everyone were committed to acting perfectly ethically with respect to each other (DR: 6:312).

Its Solution: Lawful Public Authority

The solution, Kant argues, is

> *. . . a system of laws for a people . . . which because they affect one another, need a rightful condition under a will uniting them, a constitution (constituto)*, so that they may enjoy what is laid down as right.
>
> (DR: 6:311)

Only a *united* will constituted in a system of legitimate public laws and institutions has the moral authority to define, enforce, and adjudicate our respective rights and duties in cases subject to reasonable dispute (see DR: 6:313–14). Why does the united will have such authority? Kant appeals to an idea central to the social contract tradition, consent:

> when someone makes arrangements about another, it is always possible for him to do the other wrong; but he can never do wrong in what he decides upon with regard to himself (for *volenti non fit iniuria*). Therefore only the concurring and united will of all, insofar as each decides the same thing for all and all for each, and so only the general united will of the people, can be legislative [i.e., sovereign].
>
> (DR: 6:313–14)[7]

When we agree to unite our wills with others by forming a civil state under the rule of law, we agree to be ruled by the law of that state, even in cases where we may reasonably disagree with it. Its coercive enforcement in such cases, therefore, is not wrongful. The united will settles our disputes on behalf of the parties to the dispute as well as everyone else in the political community. Kant refers to the action of the united will as "omnilateral" to distinguish it from the "unilateral" action of a private, individual will (DR: 6:256).

The only way to respect everyone's innate right of freedom under universal law in conflict cases is to unite our wills and enter a "civil condition" under the rule of law. We therefore have a duty to do so, Kant concludes:

> [Every human being] must leave the state of nature, in which each follows its own judgment, unite itself with all others (with which it cannot avoid interacting), subject itself to a public lawful external coercion, and so enter into a condition in which what is to be recognized as belonging to it is determined by law and is allotted to it by adequate power (not its own but an external power); that is, it ought above all else to enter a civil condition.
>
> (DR: 6:312)

Main Implication: The Kantian Priority of Right

The judgments of the institutions of a lawful public authority that determine our respective rights and duties in cases subject to reasonable dispute, therefore, take *moral priority* over private ethical judgments about what those rights and duties should be in such cases. I will refer to this priority as the Kantian "priority of right." To reject the judgments of legitimate public authority in such cases is to do wrong "in the highest degree," Kant says, because one rejects the basis of rightful relations with others to "hand everything over to savage violence…" (DR:6:308n).

What distinguishes Kant's political philosophy from traditional social contractarian justifications of public authority is that the priority of right is not established by reference to its good consequences. Thomas Hobbes, John Locke, and Jean-Jacques Rousseau all argue in different ways that we should respect the state's authority to settle our disputes because, otherwise, we would suffer significant negative consequences, individually and collectively.[8] But this instrumental rationale for public authority will not satisfy a deontologist like Kant who holds that doing what is morally right is qualitatively more important than doing what has the best consequences. If my duty is to do what is right *irrespective of the consequences*, and I reasonably believe that the law's judgment in some disputed case is wrong, then it seems that I should reject that judgment and do what I believe is right in the case, even if such disregard for the law's authority may eventually result in the violence or insecurity that Hobbes and Locke fear. Kant explains why respecting the priority of the law's judgment in such a case *is* the morally right thing to do irrespective of its consequences.

Kant's main insight into political theory is that the problem of reasonable disputes over rights can be solved by appealing to the moral authority of the united will as constituted in the tripartite institutions and public laws of a legitimate civil state. Only the united will can settle reasonable disputes over our rights in a way that respects the freedom and equality of everyone. Hence, we must respect the rule of law in cases of reasonable dispute over our rights, even when we believe that the law's judgment is wrong.

Now, there are indeed cases in which one's duty is to resist unjust law and stand on what is morally right. Kant would agree that when positive laws clearly violate fundamental rights of freedom and equality, then one has no duty to obey them; for example, Kant rejects a constitution establishing a caste system because it could not secure everyone's consent (T: 8:397). The duty to respect the priority of right depends on the prior duty to respect the innate right of equality. Hence when a law clearly violates the innate right of equality, the Kantian priority of right does not operate.

Kant's political philosophy thus provides a partial answer to the question of how to program autonomous machine agents to act morally rightly in cases where what is right is subject to reasonable dispute.[9] We should program them to respect the Kantian priority of

right. The behavior of an autonomous machine agent that obeys the law in order to respect the Kantian priority of right will thus sometimes diverge from a merely legal machine that obeys the law in order to minimize its legal liability. A rightful autonomous machine agent would ignore laws that violate fundamental rights of freedom and equality, whereas a merely legal machine agent would likely comply with them. Conversely, while a merely legal machine might ignore or evade legitimate laws that are unlikely to be enforced, a rightful machine would still respect them.

In the next section, I show how appeal to the Kantian priority of right resolves the conflicts of rights in the (in)famous "trolley problem" for autonomous machine agents.

The Kantian Priority of Right and the Trolley Problem

Consider the following hypothetical accident scenario ("Driver") (Foot 1967, 3): Suppose you are the driver of a trolley whose brakes have failed. The trolley is approaching a junction in the tracks. On the track ahead are five people who will be struck and killed if you maintain course, while on a side track is one person who will be killed if you turn the trolley. Should you maintain course or turn the trolley? Most people (about 85%) say they would turn the trolley (see, e.g., Mikhail 2007). Contrast this scenario ("Footbridge") (Thomson 1976, 207–208): Suppose you are standing on a footbridge overlooking the trolley's track. The five are still stranded below, but now there is no side track. Standing next to you on the footbridge is a large man. If you push him off the footbridge onto the track, then he would be struck and killed, but the collision would stop the runaway trolley, saving the five. Should you push the large man or not? Most people (about 90%) say they would *not* do so (Mikhail 2007).

The original trolley "problem" posed by Phillipa Foot is the problem of how prevailing moral intuitions in Driver can be reconciled with those in cases like Footbridge, since most people are willing to kill one person to avoid killing five in Driver but not in Footbridge (Foot 1967, 3). How can prevailing moral intuitions in Driver and in cases like Footbridge be simultaneously rational? Foot argues that the answer is that "negative" duties such as to avoid killing others are more important than "positive" duties such as to aid them (Foot 1967, 5–6).[10] In Driver, the conflict is between negative duties not to kill one and not to kill five, Foot says, and since you must, therefore, violate a negative duty not to kill regardless of what you do, it is rational to turn the trolley so as to violate the fewest negative duties (Foot 1967, 5). In Footbridge, by contrast, the conflict is between a negative duty not to kill one (the large man) and a *positive* duty to aid the five, Foot says. In such a case, the negative duty should take priority over the positive duty (Foot 1967, 6). It is therefore rational to kill one to spare five in Driver but not do so in Footbridge, according to Foot.

The Solution to the Original Trolley Problem

Judith Jarvis Thomson criticizes Foot's analysis, pointing out that Foot needs to provide some account of how and why "negative" duties to avoid acts such as killing others should take priority over "positive" duties to perform acts such as aiding others (Thomson 2008, 372). I argue that appeal to the Kantian priority of right can provide this account. Negative duties not to kill in Foot's trolley problem take moral priority not because they are negative duties but because they are *legal* duties authoritatively determined in public law, whereas positive duties to aid others in cases like Footbridge are *ethical* duties. Foot's distinction between negative and positive duties roughly correlates with Kant's distinction between legal and ethical duties, since legal duties are often negative and some important positive

ethical duties cannot be legal duties.[11] But the relevant distinction is between legal and ethical duties.

Consider Footbridge again: Suppose you are one of the 10% who believes that your ethical duty is to push the large man because that would save the most lives. But the large man's right to his life has already been authoritatively determined by public law to include at least the right not to be coerced to die in order to aid others. The Kantian priority of right, therefore, controls. Your moral duty is to defer to the legitimate determination of the law concerning the scope of the large man's right to his life, whatever your private ethical judgment in the case may be. To do otherwise is to reject the basis of rightful relations with others. The prevailing intuition that one should not push the large man in Footbridge is, therefore, rational.

Now contrast Driver: Just as you had a legal duty not to kill the large man in Footbridge by pushing him, so here in Driver you have a legal duty not to kill the one on the sidetrack by turning the trolley. But *because you are the driver,* you also have a legal duty not to kill the five on the main track by maintaining the trolley's course. As the driver of the trolley, you are subject to a legal duty of reasonable care when driving that includes at least some duty to avoid collisions that injure or kill people. To see this prior legal duty of care more clearly, suppose there is *no one* on the sidetrack.[12] Or compare an analogous case where you are the driver of a car on a multilane highway (Thomson, 2008, 369). If five people were stranded in the lane ahead, and you could safely change lanes to avoid them, then choosing to maintain course and kill them would violate a legal duty to drive with reasonable care (see Thomson, 2008, 369).[13]

These comparison cases show that the driver is subject to *some* legal duty of reasonable care with respect to the five; the question is one of its scope and shape. In cases of conflict between *legal* duties such as in Driver, the priority of right does not control, and this is what distinguishes Driver from Footbridge. The question of the scope of the driver's duty to drive safely in a case such as Driver has not been authoritatively resolved in public law. There is a reasonable legal case for holding the driver responsible for the injuries of those the trolley kills, regardless of what choice the driver makes. Since the resolution of the conflict between the driver's legal duties in Driver is unclear, it seems rational (as Foot suggests) to minimize harm.[14] The prevailing intuition to turn the trolley in Driver is thus also rational. This solves Foot's original trolley problem.[15]

The Autonomous Trolley Problem

While the analysis of the trolley problem is somewhat different for autonomous machine agents, its solution by appeal to the Kantian priority of right remains the same. In Footbridge, if the manufacturer of an autonomous robot programmed it to push the large man into the trolley's path because that would minimize lives lost, then the manufacturer would be legally liable for battery and, perhaps, murder. Hence, the manufacturer's legal duty to program the robot to avoid killing the large man in Footbridge is clear and takes priority over the ethical duty the manufacturer may have to program the robot to aid the five.

In Driver, the manufacturer is subject to legal liability regardless of what it programs the autonomous trolley to do. If the manufacturer programs the trolley to turn in Driver, then the manufacturer will be liable for battery (or murder) of the one on the sidetrack. The defense that the killing is necessary in order to avoid the greater evil of killing five will likely fail because the doctrine of legal necessity typically does not excuse intentional acts that cause bodily injury or death (Wu 2020, 9). If, however, the manufacturer programs the trolley to maintain course, then the manufacturer likely would be held

AU: Please advise whether the edit made in the sentence "In Driver, the manufacturer is subject…" retains the intended meaning.

liable for the deaths of the five on the main track on a theory of strict product liability. Survivors of the five killed would argue that the car is subject to a design defect, since a reasonable alternative design that kills one to spare five would achieve a better balance of expected utility than the (defective) design that killed five to spare one (see Wu 2020, 8). Since the manufacturer's legal duty is unclear in Driver, the priority of right does not control. It is therefore rational for the manufacturer to program the trolley to turn in Driver in order to minimize harm. This solves the trolley problem for autonomous machine agents.

Resolving Dilemmas: "Driver" Redux

Now, one might object that my appeal to minimizing harm to resolve Driver seems ad hoc, and indeed, my analysis of Driver was too quick. Let us assume, as indeed Foot and Thomson both do, that the conflict in Driver is between narrow, negative duties not to kill each of the five by maintaining course, and a narrow, negative duty not to kill the one on the sidetrack by turning the trolley. Foot argues that it is better to violate only one such negative duty not to kill rather than five, and that this is why you should turn the trolley in Driver (Foot 1967, 5).

But while minimizing violations of legal rights of the same kind seems rational, doing so may be contestable as a principle of justice. It may not be clear why it is just to allow the violation of one person's rights in order to achieve the greater good of avoiding violating five people's rights. The one whose rights are violated is wronged, regardless. The conflict in Driver therefore appears to be a genuine dilemma cast between narrow legal duties, where no matter what the driver does, she may reasonably be understood to have acted wrongly. How should agents act in dilemmas cast between two ~~legitimate~~ legal obligations? What is the role of ethical principles such as harm minimization in resolving such conflicts?

I argue that the Kantian priority of right makes two demands relevant to how agents facing such dilemmas may appeal to such principles[16]: (1) First, the agent should try to formulate the dilemma by way of some legal analysis of the duties in conflict. Only after determining that one's legal obligations are indeed in intractable conflict may the agent resort to private ethical principles or preferences to determine right action. (2) Second, any decision taken in a dilemma case must be justified by reference to a ~~reasonable~~ legal argument. Public laws that determine rights are just only if everyone can consent to them, but one cannot consent to a law that lacks any rational basis whatsoever (T: 8:297). Completely irrational laws are incompatible with the consent needed to unite our wills under the rule of law.

The moral reasoning component of an autonomous machine agent thus should (1) formulate alternative consistent sets of enforceable law applicable to its goals, and then (2) make a choice between these sets, justifying its choice by citing qualifications on applicable legal rules as necessary to form a consistent set. The choice between consistent sets of law should ~~then~~ be made by appeal to some principle of justice, if possible, but failing that, may ~~then~~ be made by some fallback ethical principle such as harm minimization. A machine that resolved dilemmas in this way would respect both the priority of public laws over private ethical theories or preferences and the demand that its decisions have some minimal rational legal justification.[17] Contrast a merely legal machine that would resolve legal conflicts by calculating which action would reduce the risk of liability or culpability. This calculation likely would be driven by a prediction as to how a court would resolve the conflict case if it were litigated. For a rightful machine, ~~however,~~ such conflict cases would be resolved by principles of justice, and failing that, principles of ethics.

Conclusion

I have argued the following four main claims:

1 Autonomous machine agents programmed to enact a popular majority's ethical preferences in controversial cases involving rights would be *immoral* machines that often act in ways that violate the moral equality of persons.
2 Autonomous machine agents must respect the moral priority of the judgments of legitimate public authority over private ethical preferences in cases where rights are subject to reasonable dispute. To act morally rightly, autonomous machine agents must respect the Kantian priority of right.
3 Appeal to the Kantian priority of right solves the original "trolley problem" by showing how prevailing intuitions that it is permissible to kill one to spare five in the Driver variation, but not to do so in cases like the Footbridge variation, are simultaneously rational. The priority of right controls in Footbridge, but not in Driver.
4 The rationale for the priority of right illuminates how to handle dilemmas in the law. Dilemmas should be resolved into competing consistent sets of applicable law, and if no further principle of justice indicates which to select, then the machine may select one by applying supplemental ethical principles or preferences.

Notes

1 The challenge is not to build machine agents that act *morally autonomously* in the Kantian sense of that term. While autonomous machine agents can be programmed to do what is right, they cannot be programmed to freely choose to do what is right for the reason that it is right, which is what Kantian moral autonomy requires (G: 4:397). In artificial intelligence contexts, machine agency consists in the ability to act on an external environment, and machine agent autonomy consists in the ability to progressively alter how the agent acts on the environment as it learns more about it (Russell and Norvig 2010). Autonomous machine agents are programmed with a set of objectives, rewards and punishments, constraints, inference rules, or other performance measures, and then programmed to learn how to act in ways that optimize or satisfy those measures across a wide range of situations and environments. Hence while autonomous machine agents might be understood to have various "incentives" for action oriented toward achieving competing performance measures, they are not capable of freely choosing for themselves which such incentives to take as their motivating reason for action. They will always act in accordance with whatever such incentives best optimize or satisfy their performance measures in the ways that they have been programmed. See Anderson and Anderson (2011) for a general introduction to machine ethics.
2 Kant distinguishes legal duties ("duties of right") from ethical duties ("duties of virtue") and argues that in controversial cases, legal duties take priority (see DV: 6: 379). I discuss the Kantian priority of right in the second section.
3 In another recent effort along the same lines ("Delphi"), a deep neural network was trained on 1.7 million human-labeled examples as well as a number of other sources in order to model "common sense morality," defined as "ethical criteria and principles to which a majority of people instinctively agree" (Jiang et al. 2021, 6). Delphi's creators then assert that autonomous machine agents should be programmed to act in accordance with some model of popular ethical preferences like that of Delphi (Jiang et al. 2021, 2–4).
4 The social contract tradition begins with Thomas Hobbes and includes John Locke, Jean Jacques Rousseau, Immanuel Kant, and in recent times, John Rawls, among others. The tradition has two strands, a Hobbesian strand in which coercive state power is justified because it is necessary to avoid the insecurity that disputes over our rights will otherwise generate in a "state of nature," and a Kantian strand in which state power under the rule of law is justified because it is necessary to meet our prior natural duty to treat each other as free and equal persons.
5 Note that the Moral Machine Experiment does not collect preferences with respect to the race, ethnicity, or LGBT status of characters in its accident scenarios. The experiment thus appears to implicitly assume that acting on preferences with respect to these attributes would be unethical, regardless of what the global majority prefers.

6 The libertarian idea that rights of freedom might be naturally self-limiting in a system of equal freedom has been subjected to decisive criticism (Hart 1973, 543, 547–550; Rawls 1993, 291–292). See Wright (2022) for criticism of an effort to revive this idea in a Kantian context.
7 Kant argues that the legislative power is sovereign over the executive and judicial powers (DR: 6:313).
8 The classic texts are Hobbes' *Leviathan*, Locke's *The Two Treatises of Civil Government*, and Rousseau's *The Social Contract*.
9 The answer is only partial because there may still be reasonable disagreement over what action is ethical in cases that do not involve rights. Principles of justice do not apply to such cases (see DR: 6:230).
10 Foot rejects the "Doctrine of Double Effect," which she says she had previously thought resolves the problem (Foot 1967, 6).
11 I do not mean to imply that all legal duties are negative; for example, one has a positive legal duty to pay one's taxes. And it is obvious that many ethical duties such as avoiding lying are negative duties (which are also sometimes legal duties such as to avoid fraud or libel). There is a rough correlation between negative duties and legal duties because justice requires that legal duties precisely specify the actions that will satisfy them. This is easier to do when the duty is negative. At the same time, many positive ethical duties to take up ends such as others' happiness or one's own perfection cannot be precisely specified.
12 This alternative reveals the flaw in the defense that the driver who fails to turn takes no "action" to cause the death of the five. This defense will reduce to the claim that the driver has no prior legal duty of reasonable care with respect to the five. If there is such a duty, then failing to perform it is what causes their deaths (i.e., it is an "action by omission"). But as the scenario where there is no one on the side track makes clear, the driver is subject to some prior duty with respect to the five.
13 In this case, if changing lanes would also kill someone, then the "sudden emergency" doctrine may shield the human driver from liability (Wu 2020, 10).
14 I discuss the relationship between the ethical principle of minimizing harm and the legal duties in the case in the next main fourth section.
15 I ignore another popular variation of the trolley problem ("Bystander") because it is a bad thought experiment. Bystander is the same as Driver, except that instead of being the driver of the trolley, you are a bystander standing next to a lever that you could pull (or not) in order to turn the trolley to the side track, so killing one and sparing five. Bystander is posed ambiguously. In Footbridge, unlike Driver, you have no prior legal duty requiring you to prevent the trolley from killing the five, because there is no general legal duty to help or protect others. But since the "bystander" in this thought experiment exercises a level of control over the trolley's operation as complete as that of a driver, it is plausible to think that the bystander might be subject to a legal duty similar to the driver's prior duty of reasonable care to operate the trolley safely. If the bystander is subject to such a prior legal duty, then the case is like Driver and there is a conflict between legal duties. If, however, the bystander is not subject to such a prior legal duty, then the case is like Footbridge and the priority of right controls. Perhaps the bystander is not subject to a duty of reasonable care here despite her control over the trolley because unlike the driver, she is presumably not employed to operate it. While control is the most important factor establishing the duty of reasonable care, other factors should be weighed as well. Rational intuitions in Bystander will thus shift according to whether subjects draw an analogy with Footbridge or with Driver. In experiments where a case like Footbridge, rather than Driver, is presented to subjects before Bystander, many fewer would choose to turn the trolley in Bystander, and those who still would are less sure about their decision to do so (see Petrinovich and O'Neill, 1996, 156–158). Such "ordering" effects confound intuitions in every variation of the trolley problem except Driver and Footbridge (Liao et al. 2007). Thought experiments where rational intuitions shift because the problem is posed ambiguously are bad thought experiments. Any conclusions drawn from them will rest on equivocation.
16 These normative demands are somewhat similar to the demands that John Rawls' doctrine of public reason makes of citizens engaged in political activity (see Rawls 1993, 212–254).
17 For some discussion of a deontic logic appropriate for handling legal dilemmas in this way, see Wright (2021, 233–235).

References

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge: Cambridge University Press.
Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., et al. (2018). The Moral Machine Experiment. *Nature* 563, 59–64.

Foot, P. (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5, 5–15.
Hart, H. L. A. (1973). Rawls on Liberty and Its Priority. *The University of Chicago Law Review* 40(3), 534–555.
Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Forbes, M., Borchardt, J., et al. (2021). Delphi: Towards Machine Ethics and Norms. *arXiv preprint* arXiv:2110.07574.
Kant, I. (1992). In P. Guyer and A. Wood (Eds.), *The Cambridge Edition of the Works of Immanuel Kant*. Cambridge: Cambridge University Press. All references to Kant's work are from the Cambridge edition unless otherwise noted. Citations are made according to standard Academy pagination.
The Doctrine of Right, Part One of The Metaphysics of Morals, trans. M. Gregor [DR]
The Doctrine of Virtue, Part Two of The Metaphysics of Morals, trans. M. Gregor [DV]
Groundwork of the Metaphysics of Morals, trans. M. Gregor [G]
On the Common Saying: 'That May Be Correct in Theory but It Is of No Use in Practice', trans. M. Gregor. [T]
Liao, S., Wiegmann, A., Alexander, J., & G. Vong. (2012). Putting the Trolley in Order: Experimental Philosophy and the Loop Case. *Philosophical Psychology* 25(5), 661–671.
Mikhail, J. (2007). Universal Moral Grammar: Theory, Evidence, and the Future. *Trends in Cognitive Sciences* 11(4), 143–152.
Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., & Procaccia, A. (2018). A Voting-Based System for Ethical Decision Making. In *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1). https://doi.org/10.1609/aaai.v32i1.11512
Petrinovich, L., & O'Neill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology* 17, 145–171.
Rawls, J. (1993). *Political Liberalism*. New York: Columbia University Press.
Thomson, J. (1976). Killing, Letting Die, and the Trolley Problem. *The Monist* 59, 204–217.
Thomson, J. (2008). Turning the Trolley. *Philosophy & Public Affairs* 36(4), 359–374.
Wright, A. (2021). Rightful Machines. In H. Kim and D. Schönecker(Eds.). *Kant and Artificial Intelligence*. Walter de Gruyter GmbH & Co KG.
Wright, A. (2022). Kantian Freedom as "Purposiveness". *Kant-Studien* 113~~ (forthcoming)~~.
Wu, S. S. (2019). Autonomous Vehicles, Trolley Problems, and the Law. *Ethics and Information Technology* 22 (1), 1–13.18