8

Discrimination in Algorithmic Trolley Problems

Derek Leben

Introduction

Some uses of the Trolley Problem are productive, while others are less so. One productive use is a testing ground for illustrating the predictions of normative theories; for example, Theory A may both pull the switch in "Bystander" and push the large man in "Footbridge," while Theory B may do neither. Another productive use is gathering evidence for theories of moral psychology; for example, if cross-cultural surveys show that people tend to judge "Loop track" as less permissible than "Switch," but more permissible than "Footbridge," then one viable psychological hypothesis is that the moral intuitions of laypeople are sensitive to the means/ends distinction. Finally, the Trolley Problem can be used as an abstraction or analogy when thinking about real-world situations in which the welfare of a small group must be sacrificed in order to prevent unavoidable harm to a larger group, such as Henry Stimson's justification for dropping atomic bombs on Hiroshima and Nagasaki, as a necessary means for preventing a full-scale invasion of mainland Japan. However, this paper is about an *unproductive* use of the Trolley Problem, namely, as a way of examining judgments about *ethically irrelevant information*, like the relative value of certain demographic groups. This paper will argue that a determination of which demographic features are relevant to the Trolley Problem is an issue that must be settled prior to any decision procedure within the scenario itself, and it is instead dependent on considerations about the task that is being performed.

The discussion in this paper will focus on the way in which the structure of a Trolley Problem appears in tasks that are being carried out by AVs and risk-assessment procedures employed by parole boards in five US states. Looking at algorithmic decision-making in Trolley Problems is important because it demonstrates the need for developing well-defined effective procedures for reasoning about these problems, rather than relying on "gut intuitions," or deferring to human character and choices. When machines are forced to make decisions about sacrificing a small group of people who were not otherwise in danger in

Derek Leben, *Discrimination in Algorithmic Trolley Problems* In: *Autonomous Vehicle Ethics*. Edited by: Ryan Jenkins, David Černý, and Tomáš Hříbek, Oxford University Press.
DOI: 10.1093/oso/9780197639191.003.0008

order to benefit a larger group from some looming harm, this quickly brings the abstract nature of the Trolley Problem down to Earth.

Autonomous vehicles (AVs) navigating through any populated environment will require unavoidable decisions about which paths are better and worse than others. Paths without collisions are obviously preferable to paths with collisions. But it would be absurd to consider all collisions as equally bad. Colliding with a group of pedestrians is worse than colliding with a cardboard box, all else being equal. It follows that AVs must be equipped with some way of discriminating between cardboard boxes and pedestrians. But should they also judge collisions with some pedestrians as worse than others? We could certainly design AVs that are programmed to discriminate between pedestrians or vehicle passengers on the basis of age, gender, socioeconomic status, race, religion, criminal background, and so on. There is good experimental evidence suggesting that most people do, in fact, use group membership as a basis for their judgments about collision rankings in trolley-style dilemmas (Awad et al. 2018). For example, the majority of people in most cultures view collisions with women as worse than collisions with men. However, this conflicts with an anti-discrimination principle endorsed by most moral theories and international statements of human rights: depriving individuals of some resources or opportunities on the basis of group membership alone is morally wrong. I will argue that setting up a Trolley Problem with these sorts of irrelevant demographic factors is a violation of anti-discrimination principles.

In models that are used by parole boards to help assess the riskiness of a prisoner, there will always be some number of errors. For parole decisions, these errors mean either denying parole to low-risk prisoners (false negatives) or allowing high-risk prisoners to go free (false positives). With automated decision-making, like Equivant's COMPAS model, it is possible to estimate and adjust these error rates, along with the predicted rise in violent crime within the larger community (Corbett-Davies et al. 2017). Thus, companies that design models like COMPAS must decide how much they are willing to increase the rate of low-risk prisoners that are mistakenly kept in jail in order to prevent a rise in violent crime within the society. Like the task of AV navigation, this has the structure of a Trolley Problem. One might initially think, like Bernard Williams (Smart and Williams 1973), that it is always morally wrong to punish an innocent person, no matter what the consequences. Yet if we set up a parole system to ensure that low-risk prisoners are never mistakenly kept in jail, this will dramatically increase the violent crime rates. One might think that this can be avoided by focusing on certain demographic features of prisoners that are statistically correlated with being high-risk, such as age, yet I will argue that the use of age in addressing this Trolley Problem is irrelevant and thus discriminatory, while it

may be relevant in addressing the AV Trolley Problem. Thus, the Trolley Problem itself cannot address what counts as discriminatory in either context.

There is a frustrating vagueness in most statements of human rights about the definition of "discrimination." On a simple standard, merely disadvantaging an individual on the basis of group membership counts as discrimination. However, I will advocate a "task-relevance" standard for discrimination, where we also care about whether group membership is *relevant* for the task at hand. Under this standard, age may be a relevant factor in algorithms for AVs, but criminal history would not. By contrast, age would not be a relevant factor to use in criminal justice algorithms, while criminal history may be relevant. Even if both cases involve the structure of a Trolley Problem (sacrificing some to save many), the question of which information is relevant or discriminatory is independent of these questions about sacrifice.

The Task-Relevance Standard

We will start by developing a standard for discriminatory practices. Under a simple "membership" standard, discrimination is any action that deprives someone of goods on the basis of their membership in some group. Yet this simple standard fails to distinguish between the following two scenarios:

Ted the Bank Manager
Ted manages a bank and is hiring some new tellers. He prefers to be around men at work, so he only contacts the male applicants for interviews.

Bill the Theater Director
Bill is directing a new production of Shakespeare's classic play, *Macbeth*. For the role of Lady Macbeth, he specifies that only female actors need apply, and only contacts the female actors.

Both of these cases involve denying people employment on the basis of group membership. However, I have the strong intuition that the first is morally wrong, but the second is acceptable. There are several logical possibilities that this intuition is consistent with. Perhaps I think all types of discrimination are wrong, but I don't define the second case as discrimination. For instance, Wasserman (1998) and Richards and Lucas (1985) suggest that the very concept of discrimination entails something unfair or wrong. Alternatively, one might count both of the above cases as discrimination, but only judge that some types of discrimination are wrong. This is a so-called value-neutral definition of discrimination, as advocated by Singer (1978). As Altman (2015) notes, it's possible to define "discrimination" in a way where it is potentially acceptable:

> We can, in fact, distinguish a moralized from a non-moralized concept of discrimination. The moralized concept picks out acts, practices or policies insofar as they *wrongfully* impose a relative disadvantage on persons based on their membership in a salient social group of a suitable sort. The non-moralized concept simply dispenses with the adverb "wrongfully."

Whether the concept of discrimination has an essential connection to a moral judgment, being what Williams (1985) calls a "thick moral concept," or only an accidental one, a "thin moral concept," is not especially important here. What is important is that some additional standard is needed to specify exactly *when* discrimination is morally wrong. If discrimination is wrong in the Bank Manager case, but not in the Theater Director case, we need an explanation of why that goes beyond just the basic membership standard. A better explanation for what is driving the differences in these cases (and what's wrong with discrimination) appeals to the nature of the task itself. Namely, gender is *relevant* to the task of playing Lady Macbeth, but *irrelevant* to the task of being a good bank teller. Call this the "task-relevant" standard for discrimination (it is also sometimes called the "irrelevance" standard). There are two arguments for adopting this standard over the simple membership standard.

First, the task-relative standard better explains people's usage of the term "discrimination." According to a simple membership standard, both count as discrimination; but this is clearly not how people are using the term. The task-relevance standard for discrimination nicely explains people's judgments about the Ted and Bill cases. Furthermore, the standard is found in the form of standard exceptions within anti-discrimination laws. Halldenius (2017) notes that the European Union anti-discrimination laws make exceptions for differential treatment when "such a characteristic constitutes a genuine and determining occupational requirement." She concludes from these exceptions that "irrelevance is part of the definition of what discrimination is." The second argument for adopting a task-relevance standard is that it better connects to a moral account of why discrimination is wrong, and thus explains the correlation with moral impermissibility. Singer (1978) points out that irrelevance is inherently arbitrary, and most moral theories reject arbitrary standards for disadvantage, whether it is because arbitrariness is disrespectful, harmful, or violates the just desert of moral patients. The arbitrariness of irrelevant treatment also may play a role in why other actions are morally wrong, even when it doesn't involve group membership. For instance, imagine that Ted the bank manager refuses to hire an applicant because that applicant has visited Madagascar twenty times. This isn't a judgment based on simple membership in a group, but if it's unacceptable behavior, then it's because frequent trips to Madagascar are irrelevant to the task of being a good bank teller. On the basis of these two arguments the rest

of this paper will now employ the task-relevance standard to determine when algorithms are indeed violating discrimination principles.

The Task-Relevance Standard for AVs

AV behavior, like human behavior, is a product of perception and decision-making processes. This paper will be focusing entirely with the decision-making side of machine behavior, while assuming that the problems of perception can be solved, at least in principle. Thus, exactly how AVs gather information about age, gender, religion, socioeconomic status, and other factors is not what we care about. Instead, the discussion is a hypothetical: if the vehicle could have access to all the information about people in its field of vision, what types of information *should* it use to make judgments?[1]

In the "Moral Machine" experiment (Awad et al. 2018), researchers from the MIT Media Lab presented participants all around the world with trolley-style dilemmas where an AV loses its braking ability and must decide between two paths. The paths differ along a range of variables which include whether the vehicle swerves or stays straight, whether people in the path are jaywalking (crossing the street without permission), and whether the path leads to collisions with obstacles or people. For our discussion, the most important features are the demographic variables about the passengers in the AV and the pedestrians in its path. The variables included the following:

Gender (male, female)
Occupation (doctor, athlete, executive, homeless)
Age (old, adult, child, stroller)
Criminal background (criminal, noncriminal)
Body type (fat, not fat)[2]

The researchers found that, in global aggregate preferences, people ranked the following combinations of features from "most saved" to "least saved":

Baby (stroller)
Boy
Girl
Pregnant woman
Male doctor
Female doctor
Female athlete
Executive female

Male athlete
Executive male
Fat woman
Fat man
Homeless
Old man
Old woman
Dog
Criminal
Cat

This experiment provides interesting data about preferences and biases. The authors are (correctly) adamant that we should not draw normative conclusions from the data. If we were to program AVs to detect these features and rank pedestrians according to the aforementioned metric, there is an obvious objection that evaluating a collision with one person as worse than another on the basis of group membership alone counts as discrimination. I take this to be an obvious problem. But the more interesting question is: does the use of *any* of these factors count as discrimination, or are there some that may be morally justified?

According to a simple membership standard, we should not care about group membership as a basis for any deprivation of goods, and all the variables must be eliminated from consideration. But employing the task-relevance standard, we *can* justifiably ask whether any of the variables have an impact on evaluating the effects of a collision on each person's health. Some features clearly do not have any relevance to this consideration, such as criminal background and occupation. Homeless people are just as likely as doctors and athletes to be injured in a car crash. However, age (and perhaps other group membership variables) will almost certainly have effects on how likely it is that a pedestrian will be seriously harmed in a collision. Very young and very old people are generally more vulnerable than average adults. If the goal of our task is to evaluate the likely harm to each pedestrian in a collision, then age is clearly a relevant factor.

This conclusion is relatively intuitive when it comes to ranking collisions with children as worse than those with adults, but it also produces a counterintuitive prediction that elderly adults should be protected more than average-aged adults. According to Awad et al., the elderly are ranked almost at the bottom of the value scale by most people, just above dogs, cats, and criminals. One response is that people's judgments are systematically mistaken here, which wouldn't be surprising. Almost everyone strongly prefers to save those who are genetically related in trolley-style dilemmas (Bleske-Rechek et al. 2010), but there is almost no way to justify the claim that one's own family is more valuable than the families

of others (much less implement it into a formal rule). Another response is to re-evaluate the way that harm is being calculated in the problem, where we are not considering the end result of a collision but rather the loss of health or potential life from one's prior state. These sorts of arguments are often made in emergency rescue situations to justify prioritizing the younger over the older, on the grounds that the elderly have "less to lose" than someone at the beginning of his or her life (Persad et al. 2009). Whether such an argument carries over to the context of AV navigation systems depends on conceptual questions about the nature of the tasks. As with the Switch and Footbridge versions of the trolley problem, one might have an intuition that there is some important difference between redirecting harms in emergency medical rescues and redirecting harms in path navigation systems, but this requires a specific explanation of exactly why they differ. Lacking such an explanation, and granting the relevance of future potential to measurements of harm, it follows that preferring collisions with the elderly to collisions with adults or children may not be unjustified discrimination.

These two considerations, vulnerability and loss of potential, are conceptual reasons why treating the very young and very old differently in AV collision evaluations may not count as unjustified discrimination. I have argued elsewhere that both of these considerations can be incorporated within a simple measurement of harm in collisions as a change in prior likelihood of survival (Leben 2018; see also Keeling 2018 for criticisms).

Imagine three paths where each results in a collision, one with a child, one with an adult, and one with an elderly person (Figure 8.1). Assign some distribution of prior likelihoods of survival to each, perhaps something like A = (.99), B = (.99), and C = (.80). Now consider that a collision under the same conditions would result in larger losses to the very old and very young, because of their greater vulnerability. Say that the estimated survival rates for each person in these collisions (based on previous data about similar collisions with these age groups) are A = (.65), B = (.80), and C = (.50), reflecting the greater impact on very young and very old pedestrians. If we were only looking to minimize the worst outcomes, the paths would be ranked as B > A > C. If we were seeking to minimize the worst losses from previous states, then the paths would be ranked as B > C > A. There are obviously other ranking principles that could be used, such as the loss in percentage of prior survival likelihood. I am not going to adjudicate between different ranking principles here; the purpose of this discussion is only to note that none of these ranking principles are discriminatory, assuming they are generated on the basis of information that is relevant to evaluating harms.

If the function of an AV navigation algorithm is to evaluate collisions based on their predicted health outcomes, then some features are relevant to that task and others are irrelevant. Once we've established the core function of the algorithm,

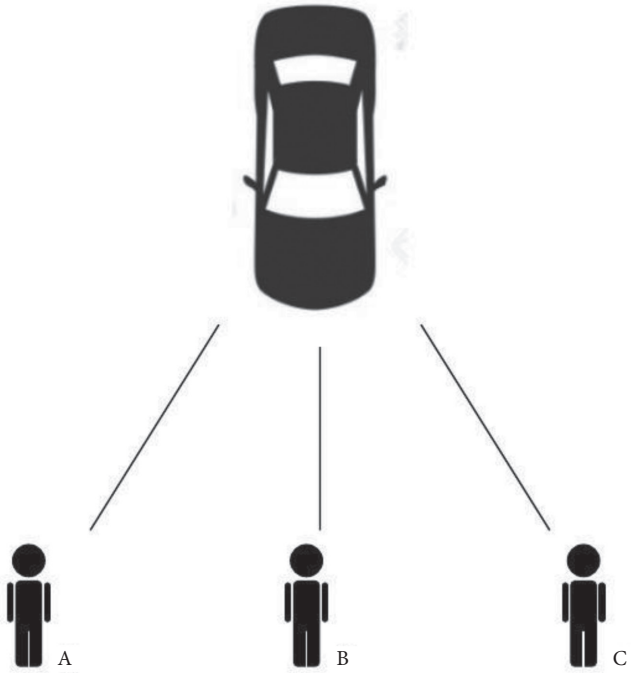

Figure 8.1 Three potential paths for an autonomous vehicle, resulting in a collision with (A) a child, (B) an adult, or (C) an elderly adult.

it now becomes an empirical question which features will have an impact on accomplishing that function. I am assuming that age will be a statistical predictor of health outcomes in collisions, while occupation is not, but this is a question that can (and should) be settled with data. If the effects of age in collisions are negligible, this is good reason for concluding that this factor is irrelevant for AVs.

Perhaps the most surprising prediction of my position is that individual history should not be taken into account by AVs. Specifically, those who are jaywalking or otherwise being negligent should not be valued less than those who are following the rules of the road. In Awad et al.'s data, a large majority of the aggregate data strongly favors those who are following the rules as opposed to those who are crossing when they shouldn't be. In my own anecdotal experience, I've found many students respond to the footbridge version of the trolley problem by insisting that the workers on the track "shouldn't have been there," while the large man on the footbridge was not doing anything negligent (usually this can be remedied by changing the scenario such that they are now kidnapped and tied to the tracks by a madman). However, if the task of the system is only to evaluate paths based on likelihood and harm, rather than desert and culpability,

then these features are morally irrelevant. On the other hand, there are contexts where trade-offs in trolley-style dilemmas *do* involve considerations of desert, and the next section will consider how criminal justice algorithms change which factors are relevant.

The Task-Relative Standard for Criminal Justice Algorithms

Assuming both AV navigation systems and criminal justice algorithms have the structure of a Trolley Problem, we might infer that if the use of age as a deciding factor is acceptable in one, then it would also be acceptable in the other. However, this is not the case. Indeed, there is a much stricter standard that we should apply within the criminal justice system, because the tasks of navigation and justice are different in kind. In US law, there is a distinction between "statistical evidence" and "individualized evidence," where the former is never acceptable in a decision procedure. A standard case study for illustrating this distinction is based on a 1945 case before the Massachusetts Supreme Court, "Smith v. Rapid Transit Inc." (317 Mass. 469, 470, 58 N.E 2d 754, 755 (1945)). The simplified version of the case looks like the following:

> **Red Bus, Blue Bus:** Smith was run off the road by a bus late at night, causing her to collide with a parked car. There are two bus companies, the Red Bus Company and Blue Bus Company. Neither Smith (nor anyone else) saw which type of bus was responsible, but 90 percent of the busses on that road are from the Red Company.

The development of algorithms (like COMPAS) which are used for assisting with decisions about bail, sentencing, and parole have brought this question into the forefront of debates about acceptable use of technology in the legal context. As a Pro Publica investigation revealed, age appears to be one of the most important factors in determining whether the defendant poses a high risk. But does this count as discrimination? One might argue that eighteen- to twenty-five-year-olds commit a disproportionate number of violent crimes, so it is reasonable to evaluate young people as a higher threat for bail, sentencing, and parole decisions. Littlejohn (2017) distinguishes two premises in this argument, one epistemic and the other moral:

> **Epistemic**: If there is sufficient statistical evidence to associate *x* with a crime, it is rational to believe that *x* is responsible for that crime.
> **Moral:** It is justified to punish *x* for the crime, on the basis of the epistemic claim.

There is something intuitively appealing about the epistemic claim. Imagine we have two urns filled with a thousand balls each, which are either red or blue. From Urn A, we select 100 balls, and 95 of them are blue. From Urn B, we select 100 balls, and 95 of them are red. It seems rational to conclude that the next ball from Urn A is likely to be blue, and the next ball from Urn B is likely to be red. Being a member of a group serves as a good statistical predictor of an individual's likely traits. Despite the apparent strength of the epistemic claim, in 1945 the Massachusetts Supreme Court famously rejected the use of statistical evidence in evaluations of legal culpability. So what's going wrong here?

Perhaps the reason why many people reject the use of statistical evidence is not because of the epistemic claim, but rather, the moral one. As Colyvan et al (2001) describe, in order to hold people morally or legally responsible, "We require more evidence than simply their membership in the reference class in question." This is related to public attitudes about police profiling. Even if the majority of violent crimes in a society are committed by a certain ethnic or religious group, many people insist that there is something wrong with police focusing more attention on members of that group without any other evidence. Those who oppose the moral claim, like Thomson (1986), Colyvan et al. (2001), and Littlejohn (2017), often do so by rejecting the epistemic claim.[3] This is possible; perhaps statistical evidence provides some evidence to believe that *x* is responsible, but that evidence is not sufficient for punishment. However, I propose an even bolder claim: even if it is rational to believe that *x* is responsible for a crime on the basis of statistical evidence, it would be discriminatory to punish *x* for the crime on the basis of that evidence alone.

Before considering the argument, let's first note that there is precedent for possessing evidence that may lead us to believe *x* is responsible, yet still discarding that evidence as morally unacceptable (i.e., accepting the epistemic relevance of some set of evidence, but rejecting it on moral grounds). In the US criminal justice system, evidence obtained without following due process, such as a confession obtained through force, is considered inadmissible in trial. This is true even when such evidence may provide compelling reasons to believe that *x* is responsible for the crime. I am arguing that discriminatory evidence has this kind of status; it may lead to conclusions that are likely to be true, but there is a moral reason not to make use of it in this context.

The moral argument against using statistical evidence in criminal justice proceedings makes use of the task-relevant discrimination principle. Everyone agrees that the use of statistical evidence is irrelevant in determining criminal guilt. Criminal justice algorithms like COMPAS are explicitly designed to merely "assist" the process of determining bail, sentencing, or parole, which are intended

to be distinct from the question of criminal culpability. However, it is impossible to distinguish the effects of these two actions; keeping people in prison when they are already there is the same as putting people in prison who are not yet there. If we agree that it is wrong to use statistical evidence for the latter, then it must also be wrong to use statistical evidence in service of the former. They both involve detaining people against their will, which is a kind of culpability where only past behaviors and mental states are relevant as evidence.

The core of this argument is an assumption about the essential function of the task of the criminal justice system. I am assuming that the task of the system involves retribution, compensation, and deterrence on the basis of individual responsibility. One could, of course, object by presenting alternative proposals for the essential function of the criminal justice system, where individual responsibility is not the most important feature of criminal guilt (this would also involve endorsing the repugnant conclusion about arresting people on the basis of mere statistical evidence). However, this is still within the bounds of my overall proposal that discrimination is to be determined by evaluating the essential function of a system. We are now merely arguing about what that essential function should be. I view this as significant progress; those who disagree about whether the use of group membership counts as discrimination may actually be in disagreement about a more fundamental question about the goals of an institution.

Conclusion

I've proposed that both AV navigation algorithms and parole algorithms must perform some tasks that have the structure of a Trolley Problem, but this does not mean that the same standards for relevant and irrelevant features will apply to both. Instead, the standards for what counts as discrimination in both cases will depend on the nature of the task, and this must be determined *outside* of decision-making procedures for the Trolley Problem. With AV navigation algorithms, the trade-offs involve mere harm alone, and so the acceptable features are those which will have some impact on likelihood of harm. For parole algorithms, however, the trade-offs involve deprivations based on desert, and thus the relevant features must be those that stem from an agent's particular behavioral history, beliefs, and desires. Setting up a Trolley Problem with irrelevant information can be dangerous and misleading, and it is extremely important to settle these questions about discrimination prior to establishing any decision procedures for a Trolley Problem scenario.

Notes

1. Clearly, if it's impossible for AVs to have access to such information, this makes the present discussion unnecessary. Yet this is not an absurd assumption, given the advances in perceptual skills that allow vehicle cameras to produce reliable estimates about pedestrians, along with the prospects of networked vehicles that can quickly identify demographic information about passengers in other vehicles, or even network with pedestrians through their personalized devices (mobile phones, smart watches, etc.).
2. Awad et al. use the term "large," but I will use the term "fat," with no necessary pejorative associations.
3. Littlejohn (2017) and Smith (2018) argue that the epistemic claim is missing something important. I agree with their suspicions, and think it's fruitful to draw a potential distinction between "evidence for the likelihood of *x*" from "reason to believe *x*." For instance, Bostrom's (2003) Simulation Argument proposes the following: if simulating conscious beings is possible, and there are many intelligent beings in the universe, it is statistically likely that we are one of the simulated ones. My hypothesis is that there's a difference between that claim being likely and having positive reason to believe it. I suspect that this is also relevant to many people's dismissal of the Simulation Argument, where it may be likely, but we lack specific reasons to believe it. However, since the target of this paper is the moral claim, I will simply grant the truth of the epistemic claim.

References

Altman, A. 2015. "Discrimination." In *The Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Winter 2016 edition. https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., and Rahwan, I. 2018. "The Moral Machine Experiment." *Nature* 563: 59–64.

Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W., and Brandt, S. J. 2010. "Evolution and the Trolley Problem: People Save Five over One Unless the One Is Young, Genetically Related, or a Romantic Partner." *Journal of Social, Evolutionary, and Cultural Psychology* 4, no. 3: 115–27.

Bostrom, N. 2003. "Are We Living in a Computer Simulation?" *Philosophical Quarterly* 53: 243–55.

Colyvan, M., Ferson, S., and Regan, H. 2001. "Is It a Crime to Belong to a Reference Class?" *The Journal of Political Philosophy* 9: 168–81.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. 2017. "Algorithmic Decision-Making and the Cost of Fairness." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery*.

Halldenius, L. 2017. "Discrimination and Irrelevance." In *The Routledge Handbook of Discrimination*, edited by Kasper Lippert-Rasmussen, 108–18. New York: Routledge.

Keeling, G. 2018. "Against Leben's Rawlsian Collision Algorithm for Autonomous Vehicles." In *Philosophy and Theory of Artificial Intelligence 2017*, edited by V. Mueller,

259–72. PT-AI 2017. Studies in Applied Philosophy, Epistemology, and Rational Ethics, vol. 44. Berlin: Springer.

Leben, D. 2018. *Ethics for Robots: How to Design a Moral Algorithm*. New York: Routledge.

Littlejohn, C. 2017. "Truth, Knowledge, and the Standard of Proof in Criminal Law." *Synthese*. 5253–86.

Persad, G., Wertheimer, A., and Emanuel, E. 2009. "Principles for Allocation of Scarce Medical Resources." *The Lancet* 373: 423–31.

Richards, J. R., and Lucas, J. R. 1985. "Discrimination." *Proceedings of the Aristotelian Society, Supplementary Volumes* 59: 53–83.

Singer, P. 1978. "Is Racial Discrimination Arbitrary?" *Philosophia* 8: 185–203.

Smart, J. J. C., and Williams, B. 1973. *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Smith, M. 2018. "When Does Evidence Suffice for Conviction? *Mind* 127: 1193–218.

Thomson, J. 1986. "Liability and Individualized Evidence." *Law and Contemporary Problems* 49: 199–219.

Wasserman, D. 1998. "The Concept of Discrimination." In *Encyclopedia of Applied Ethics*, edited by Ruth Chadwick, 805–14. San Diego, CA: Academic Press.

Williams, B. 1985. *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.