4

Trolleys and Autonomous Vehicles

New Foundations for the Ethics of Machine Learning

Jeff Behrends and John Basl

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

Introduction

Suppose that Annabelle is a member of an engineering team responsible for designing the systems that will determine the driving behavior of her company's autonomous vehicle (AV) model. She knows that the AV may at some point be in the kind of scenario now frequently discussed in ethics classrooms and popular media: no matter how the car behaves, a collision is unavoidable, risking harm to its occupants or others. We will refer to such situations as *accident scenarios*. Annabelle wants to know not which of her engineering options is in the best interest of her company, but what the morally permissible options are. She remembers studying trolley cases in her introductory philosophy class and has since read magazine and newspaper articles in which authors have likened her choice situation to that faced by the subjects in those cases. Could attending to trolley cases help Annabelle understand what she and her colleagues ought to do with respect to programming for accident scenarios?

The idea that trolley cases may in some important way be related to AV accident scenarios is probably a familiar one to most readers. The two have been discussed together not only in the philosophical and wider academic literatures but also in several popular news outlets.[1] The great majority of these discussions have taken the comparison between accident scenarios and trolley cases to be apt; to one degree or another, they approve of the idea that trolley cases are especially useful in addressing questions of the sort that confront Annabelle. We believe that this approval is, at best, unjustified. In this paper, we argue that despite genuine similarities between accident scenarios and trolley cases, verdicts about the latter do not directly inform questions about the former.

As the fictional Annabelle illustrates, we are interested in investigating a question that concerns what individual engineers, or perhaps teams of engineers, ought to do. Understandably, one might think that this is the wrong kind of question with which to occupy ourselves, so far as the ethics of AVs is concerned.[2] After all, engineers are hardly the only actors who will play a role in determining

Jeff Behrends and John Basl, *Trolleys and Autonomous Vehicles* In: *Autonomous Vehicle Ethics*. Edited by: Ryan Jenkins, David Černý, and Tomáš Hříbek, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197639191.003.0004

what the world of automated driving looks like. In particular, one might think that it is *policymakers* and their choices that are most deserving of our attention, given how significant the implications of their decisions might be. We, of course, do not deny that there are important questions to be addressed at the policy level.[3] But we do think it would be a mistake to overlook Annabelle's context entirely. This is because current policy does not constrain the options of car manufacturers in a significant way and is unlikely to do so soon, and because the choices made by engineers like Annabelle are themselves morally significant.[4] Perhaps there will be a time at which the central question we address is moot because there will simply be no genuine *options* for engineers to choose among, due to tight regulation. But that time appears far off, and we should be careful to govern our conduct appropriately in the meantime. Further, because the upshots of our position generalize to other automated computing applications, even if we are mistaken in focusing our attention on engineers like Annabelle where AVs are concerned, the most significant contributions of our argument would stand.

The paper proceeds as follows. In the second section, we characterize in more detail and attempt to motivate the view according to which trolley cases are especially relevant to choice situations like Annabelle's. In the third section, we argue that even if accident scenarios resemble trolley cases in various ways, it does not follow that verdicts about trolley cases immediately provide *any* guidance for designers like Annabelle. Roughly, that is because the engineering techniques that govern AV driving behavior introduce challenges as yet unrecognized by ethicists for guaranteeing any particular driving pattern in an accident scenario. In the final two sections, we first consider and respond to objections, and then conclude by elaborating on the implications of our work. We argue that our approach helps to illuminate the ways in which narrowly focusing on an ethical evaluation of the *behavior* of automated technologies can obscure the importance of ethical theorizing about the design choices that ultimately give rise to it.

Trolley Optimism

An unoccupied autonomous car is traveling toward a crosswalk at a relatively high speed. Suddenly, five pedestrians, tightly bunched together, enter the crosswalk; the car had not detected them earlier because they were blocked from its sensors by an obstruction on the sidewalk. While there is enough time and room for the vehicle to swerve to avoid them, doing so would put the car on course to hit a single pedestrian who has entered the crosswalk from the opposite side. The physics of the scenario allow for no other outcome: either the vehicle will continue on course, striking the five, or else it will divert, striking the one.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

It is no coincidence, of course, that the accident scenario just described bears a striking resemblance to the most commonly discussed trolley case, in which a runaway trolley car is on course to strike five workers on the track. While the trolley can be diverted onto a sidetrack, thus averting the deaths of the five, this would result in the death of a single worker standing on the second track. The parallels between the two cases are clear, and while it may take a little more imagination to accomplish, it is not so difficult to see how other accident scenarios might mirror the many varieties of trolley cases that populate the literature in normative ethics. The question that confronts us is whether anything of import hangs on the similarities between possible accident scenarios and trolley cases.

Trolley Optimists answer in the affirmative. More precisely, Trolley Optimism is a view characterized by its subscription to the following argument[5]:

1. Some accident scenarios are structurally similar to trolley cases.
2. If that is so, then engineers should work to program AVs such that they behave in accident scenarios in ways that conform to the appropriate verdicts about what should be done in the corresponding trolley cases.
3. Therefore, engineers should work to program AVs such that they behave in accident scenarios in ways that conform to the appropriate verdicts about what should be done in the corresponding trolley cases.

The Optimist's Argument, as we will refer to it, requires several points of clarification. First, Trolley Optimism is so called because it views the ethical question confronting Annabelle as answerable by a *different* ethical question, one to which philosophers have already devoted a significant amount of attention. The Trolley Optimist assures us that what appears to be a new, difficult problem arising from sophisticated technology is, at bottom, not really a new problem at all. The optimism that characterizes the view is not optimism that the problem is *easy*, but rather that the problem is not distinct; we can draw on currently available and well-researched resources to solve it, according to optimists.

Two further clarifications correspond to each of the premises. What does it mean for accident scenarios to be "structurally similar to trolley cases," as Premise 1 asserts that they are? As a rough characterization, to be revisited in the fourth section, let us for now say that an accident scenario and a trolley case are structurally similar so long as both situations involve the same number of possible routes for the vehicles involved, with each corresponding route having similar morally relevant consequences, and with similar possible causal chains for effecting those consequences. For example, consider again the accident scenario introduced above alongside the standard trolley case. In each, a vehicle is presently on course to strike five innocents. There is exactly one alternative route available for each vehicle; if that alternative is taken, harm to the

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

five will be avoided but also imposed on one innocent; and that imposition is a result of diverting the vehicle such that it will directly strike the one. We can imagine other accident scenarios which parallel other notable trolley cases in similar ways.[6]

A second clarification concerns the other premise of The Optimist's Argument. The second premise is meant to capture the optimist's idea that we can use our judgments about trolley cases to settle moral questions about how to engineer AVs. That idea admits of some flexibility; accordingly, so too does Premise 2. In particular, the optimist has options for specifying what it means to "program AVs such that they behave in accident scenarios in ways that conform to the appropriate verdicts about what should be done in the corresponding trolley cases." The most natural option, and the most widespread in the extant literature, we call *Standard Trolley Optimism*. On the standard view, conformity with the appropriate verdicts is satisfied so long as the *autonomous system behaves only in ways that would be permissible for an agent to cause a trolley to move in a structurally similar trolley case*. The Standard Trolley Optimist would have us attend to whatever the *correct* verdicts are in trolley cases and program our AVs accordingly.

Quite a different approach, though, also attempts to answer Annabelle's question by appeal to trolley cases, and we mean to characterize it, too, as an instance of Trolley Optimism. According to *Democratic Trolley Optimism*, we should attend not to the verdicts issued by moral philosophy about trolley cases, but instead to the judgments of the people. On this approach, conformity with appropriate verdicts about trolley cases is satisfied so long as *the autonomous system behaves only in ways that do not violate group preferences about how it is best for trolleys to be moved in structurally similar trolley cases*. To attach something of a rough tagline to the view, the Democratic Trolley Optimist would have us settle questions like Annabelle's *by vote*.[7]

Democratic Trolley Optimism is an interesting alternative to its standard cousin. It is also implausible, partially for reasons that are independent of the objection to Trolley Optimism that we advance below. We highlight the democratic approach here primarily to emphasize that Premise 2 is crafted in such a way as to allow the argument to capture the reasoning of a variety of specifications of Trolley Optimism. For the purposes of our argumentative aims here, the possibility of Democratic Trolley Optimism in particular is unimportant, and so we will not elaborate on the details regarding how it could and has so far been developed in the literature. The reader may find Standard Trolley Optimism easier to hold in mind as we proceed, and so we gently recommend that they do so. However, though we set aside any specific mention of Democratic Trolley Optimism in what follows, our challenge to The Optimist's Argument does not depend on which particular form of optimism one prefers.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

The Machine Learning Basis for Trolley Pessimism

Trolley Optimism is characterized simply by adoption of The Optimist's Argument. Correspondingly, we take *Trolley Pessimism* to be characterized by the argument's rejection; one is a Trolley Pessimist so long as one thinks that the argument is unsound. Note that this allows that there are pessimists who nevertheless *endorse* the conclusion. This may be surprising, but, we think, is nonetheless fitting. What is remarkable about Trolley Optimism is the way in which it *infers* a conclusion about how to program AVs *on the basis of* their putative similarities with the trolleys of moral philosophy. As we will explain in further detail below, we want to ultimately remain neutral on the claim that we ought to program AVs so that their behavior conforms with verdicts about trolley cases. *If* we should do so, though, we will need to secure that position via argumentation unlike that of the Trolley Optimist's.

We are not the first Trolley Pessimists; a number of extant challenges to the use of trolley cases in thinking about the ethics of AVs can be understood as objections to The Optimist's Argument.[8] However, we think that both optimists and pessimists have failed to recognize important features of the engineering techniques that underwrite the operations of AVs, and therefore have failed to consider a serious technological case against Trolley Optimism. Because the techniques in question govern much more than the driving of AVs, our challenge to Trolley Optimism has broad implications for the ethics of developing autonomous systems more generally. On our view, a clear understanding of how designers like Annabelle will make use of engineering techniques to create functional AVs undermines the claim that trolley cases can directly inform her decisions about the behavior of AVs in particular cases. This is because there are ethical questions about the available engineering options that precede and thereby screen off questions about how Annabelle should opt to guide the behavior of AVs in accident scenarios. Making the case for that claim requires that we spend some time attending to the engineering techniques that we have in mind.

Machine Learning and Autonomous Vehicles

Computers operate in part through the use of *algorithms*: instructions for carrying out a task, or series of tasks, given some input. In traditional programming, software engineers specify how each task is to be managed, what series of steps is to be performed on any given input to generate an output. For example, an early exercise in computer programming courses might require students to write a program that will sort a series of input numbers from highest to lowest. Prior to actually writing the code, the students must settle on a method, on a series of

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

instructions to capture in their code, that begins with an input and eventually yields the output of an appropriately sorted list. Those instructions constitute an algorithm for carrying out the assigned task.

Autonomous vehicles depend heavily on *machine learning algorithms*, which are importantly different from traditional algorithms. Whereas engineers make use of traditional algorithms to map the route from inputs to outputs by hand, so to speak, machine learning algorithms are used to *generate new algorithmic instructions not explicitly provided by a programmer*.[9]

In contrast to a traditional algorithm, a machine learning algorithm or machine learner is an algorithm that itself creates algorithms or mappings from input to output on the basis of supplied data. Instead of having a human decide how to instruct a computer to carry out some task, a programmer instead supplies a machine learning algorithm with a set of data—a *training set*. The training set provides examples of the kinds of input-output pairings that a programmer or developer desires. The machine learner generates an algorithm for moving from inputs to outputs that extrapolates from the training set to a broader range of cases, (hopefully) producing input-output pairings that are desirable for inputs not included in the training set.[10]

Let us consider a case from outside computer engineering to illustrate. A fair analogy for how some machine learning algorithms work is operant conditioning. When trying to train a dog to sit on command, a trainer might start by taking a treat, putting it in front of their dog's nose, and moving it back toward the dog's head, causing the dog to shift its weight such that it naturally sits. The trainer then rewards the dog with the treat. After several repetitions, the dog starts to sit as soon as the trainer moves the treat toward its nose. Eventually, the trainer starts uttering "Sit" during this process. Eventually, the trainer can simply say "Sit" and the dog sits and is rewarded. In this case, the dog is developing its own internal algorithm in light of the trainer rewarding certain input-output pairs.

Many algorithms are "trained" in the same way as the dog in the above example. A machine learning algorithm is provided with data about desirable input-output pairs for some set of circumstances or is rewarded on the basis of or to the extent that it generates desirable outputs for some set of inputs. The hope is that with enough data to train on, the algorithm will itself create an algorithm that takes new inputs and generates desirable outputs.

At first glance, using machine learning might seem an inefficient way to generate algorithms. If we could directly wire up the brain of a dog like we design simple computer programs, we could get a dog to sit without all the training and without concern about whether the dog will pick up on the wrong cues or remain responsive to commands in new contexts. As clever as computer engineers are at designing traditional algorithms, though, many problems are

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

so complex that it would be extremely difficult or even impossible to create a traditional algorithm to map inputs to outputs with even a fraction of the reliability of a machine learner that was trained using a large amount of data. For example, consider the problem of spam filtering. Imagine trying to specify an algorithm that reliably detects spam email messages. What would the rules look like? How would we ensure that the wide variety of spam is captured by the filter without preventing important messages from reaching our inboxes? Instead of trying to solve that problem using traditional programming techniques, a machine learner is given a huge training set of spam and non-spam emails, labeled as such. Upon review of the features of the emails in the training set, it generates an algorithm and puts it to work. That algorithm receives reinforcement every time we mark an email as "spam" or "not spam," and over time becomes better and better at allowing genuine messages into our inboxes while diverting spam elsewhere.

Autonomous driving is another case in which the traditional approach is not feasible; it makes spam filtering look like a simple task. We leave it as an exercise to the reader to try in advance to specify *precisely* the information and rules that would be required to safely and reliably get them to and from their closest grocery store, taking into account not only the other drivers on the road, temporary obstacles that arise due to weather, and so on but also, especially, the rules governing how the system translates raw sensor data into useful classifications of objects. In contrast to this approach, imagine instead that a huge number of traditional vehicles are equipped with sensors that record data about the car. This data includes, for example, global positioning system (GPS) information, data about the local environment outside of the car, and data about how the driver is responding inside the car. This data set constitutes a training set, including such information as that when the sensor data at a location are thus and so drivers output behaviors of such and such a type. A machine learning system can generate an algorithm on the basis of this data to address circumstances outside the training set.

This is why machine learning is crucial to all self-driving technologies in development today. While programmers cannot encode by hand a system architecture that will generate safe and reliable driving behavior, they can provide a machine learner with sufficient data for it to generate an algorithm or series of algorithms. Doing so may not even require thinking about how humans think about or process the rules of the road, as the previous illustration may have suggested. Similar to the spam filtering case, one way to train a self-driving car is to simply let it drive (or let it simulate driving) and when its behavior deviates from the desired behavior, correct it, thereby updating the algorithm that is used in the future.[11]

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

Machine Learning and Trolley Pessimism

The details about how machine learning works and the role that it plays in self-driving systems have important implications for thinking about the ethics of programming AVs and ground a case against Trolley Optimism.[12] To see how, first observe that it is a mistake to imagine that those developing AVs must or are even likely to be in a position to program, in the traditional sense, their vehicles to respond in particular accident scenarios in any particular way. Much of the portrayal of these issues suggests that programmers *must* make decisions about how AVs will behave in discrete accident scenarios, explicitly programming them to behave that way, and thus, programmers should think carefully about what the right thing to do is in various accident scenarios. But this misconstrues how the behavior of AVs is actually programmed. It is not the case that engineers, as Patrick Lin writes of them, "need to instruct an automated car on how to act for the entire range of foreseeable scenarios, as well as lay down guiding principles for unforeseen scenarios."[13]

This observation alone does not show the second premise of the The Optimist's Argument to be false. However, it suggests to us that Trolley Optimism has seemed plausible to some because they are imagining designing AVs within the traditional algorithm paradigm. Whereas it might seem irresponsible or negligent to engineer AVs using traditional algorithms and fail to program explicitly for all of the foreseeable accident scenarios in which an AV might find itself, it appears decidedly less so when we hold in mind the details of the machine learning paradigm. Machine learning enables certain technologies precisely because it allows us to avoid having to explicitly program for every eventuality that we care about or recognize as salient.

The centrality of machine learning to AVs does more than provide a reason to question a potential motivation for endorsing Trolley Optimism; it also provides a starting point for rejecting the second premise of The Optimist's Argument outright. This is because the machine learning paradigm for programming AV behavior forces us to think about the behavior of AVs in accident scenarios as only one part of a largely unified whole concerning the overall driving performance of the cars. How AVs behave in any particular scenario is partly a function of a wide range of decisions made about how to train up the algorithms that guide their driving. To put it another way, to decide that AVs really should be trained to respond in particular ways in certain accident scenarios presupposes answers to myriad other questions. Once we come to understand these questions and what it might mean to answer them appropriately, it becomes clear that a structural symmetry between some accident scenarios and some trolley cases does not provide sufficient reason to try to ensure that AVs mirror appropriate behavior in trolley cases.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

To illustrate this idea, it is useful to distinguish a variety of questions we might ask about the relationship among accident scenarios, trolley cases, and machine learning techniques:

Regime: How much of the training regime—the total set and arrangement of machine learning tools utilized to govern a vehicle's behavior—should be dedicated to dealing explicitly with accident scenarios?

Structure: Assuming accident scenarios are included in a training set of desirable driving maneuvers, should those scenarios be structured like trolley cases (and, if so, which ones)?

Verdict: Should verdicts about trolley cases be used to determine how to reward or punish AV driving behavior?

Behavior: Should AVs be programmed to behave in accident scenarios in ways that conform with the appropriate verdict in a similarly structured trolley case?

Trolley Optimists are focused on the question of Behavior but fail to recognize the fact that the driving behavior of AVs in any scenario is sensitive to how we answer Regime, Structure, and Verdict. For example, if the correct answer to Regime is that none of the training regime should be dedicated to accident scenarios, then the answer to Structure, Verdict, and Behavior is "No"; the fact that AVs might be in accident scenarios that are structurally similar to trolley cases does not carry the implication that we should explicitly aim to have them behave in any particular way.

To see why one might think that *none* of the training regime used to design AVs should be dedicated to accident scenarios, consider that training an AV to reliably exhibit a particular type of behavior in accident scenarios might require a training set that includes a huge number of accident scenarios. A single instance will not be enough to ensure any particular behavior, and we do not simply want the car to, for example, veer to save five lives at the cost of one; we presumably want the car to veer to save *six* lives at the cost of one, or six lives at the cost of *two*, and so on. Consider again what it takes to teach a dog to sit on command. Trainers know that dogs can pick up on many different contextual cues during training. For example, perhaps every time the trainer says "Sit," the trainer first lifts her chin; in that case, a dog may notice and respond to this subtle movement rather than the verbal command. This is the reason that professional dog trainers are careful to train in a wide range of environments and with variation in the way that verbal commands are issued. Doing so ensures that the dog is developing a reliable mapping of the desired input-output pairs. Similarly, if we want to ensure that AV behavior in accident scenarios mirrors verdicts about trolley cases, we might need to dedicate significant resources to training for such

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

scenarios. Allocating resources in that way may impose costs that we should not be willing to pay.

It is possible that including enough training data about accident scenarios to ensure some preferred behavior in them has a feedback effect on how the system responds in other driving scenarios. It is possible, for example, that by training AVs to ensure that they maximize lives saved (or whatever) in accident scenarios, the resulting system ends up in accident scenarios more frequently than it otherwise would, or it performs less safely in typical, nonaccident driving scenarios. To the extent that this is possible or likely, it provides those like Annabelle with a reason to deemphasize training for specific and presumably rare accident scenarios.

However, the case against Trolley Optimism we are advancing does not depend on our answering Regime in any particular way or on the existence of any particular tradeoffs, nor do we want to rule out that some portion of the training regime *should* be dedicated to accident scenarios. What is important for our purpose is that the potential for trade-offs like those described makes vivid that Regime is itself an ethical question: since the behavior of AVs is a function of the choices we make about the training regime, we must think about the appropriate way to constitute training sets and select training methods.[14] This ethical question is one to which we can and should apply the traditional tools of ethical theorizing, including, perhaps, trolley cases. The relevant lesson to draw from trolley cases for Regime might be, for example, that when deciding on a training regime, Annabelle ought to take seriously the moral salience of the total number of lives that might be lost depending on how a vehicle behaves across a wide range of contexts.[15] If minimizing deaths is relevant to the choice between two differently composed training sets, then trolley cases will have helped inform our advice to Annabelle.

Despite what has been said so far, we recognize that it is tempting to acknowledge that questions like Regime require ethical reflection while nevertheless insisting that our answer must be largely informed by what we think is acceptable driving behavior in particular cases, including accident scenarios. If we answer Regime in such a way that it leads to an AV behaving such that it kills the five at the cost of the one, for example, are we not therefore in a good position to conclude that we have not appropriately settled how to design the training regime?

No, we are not. To illustrate, let us assume that after highly complex ethical theorizing involving trolley cases, we are ultimately convinced of a relatively simple moral principle: it is always obligatory to act so as to minimize the number of people who die. On the surface, it may appear that this principle counsels us to abandon any machine learning regime that results in an AV hitting five when it could have instead hit one. On the contrary, though, this principle might actually

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

entail that we are *required* to enact such a regime, so long as doing so minimizes deaths across all of the driving contexts in which the AV operates.

The lesson we are after is that since there are ethical questions that precede and are in principle answerable independently of Behavior, the fact that trolley cases are structurally similar to accident scenarios does not by itself provide any reason to program AVs to behave in ways that mirror appropriate verdicts in those cases. If the answer to Behavior is "Yes," this will be due to the reasons we have for adopting particular answers to Regime, Structure, and Verdict, answers that are not provided on the basis of some structural similarity between trolley cases and AV behavior. What is needed is careful moral theorizing about whether the *explanations* for the correct verdicts in trolley cases do or do not apply to the particular decision points at which engineers like Annabelle find themselves, rather than a dispute about whether the driving options that confront an AV do or do not mirror the available routes for a trolley.[16]

Objections and Replies

The objection that we have pressed against The Optimist's Argument is a moral argument that depends on technological details about machine learning. That structure suggests two ways in which we may have erred: we could be mistaken about the technology, or we could be mistaken about the moral philosophy. We hope to alleviate concerns about both in this section.

Technological Objections

We said in the previous section that the use of machine learning in AV design raises the possibility that arranging a training regime to optimize for performance in accident scenarios may come at a cost of some moral significance elsewhere. However, one might object that we have failed to appreciate the technological possibilities afforded by machine learning or the ingenuity of computer engineers. If so, perhaps it really is possible to more or less guarantee specific driving outcomes in accident scenarios without significant moral cost.

One might try to substantiate this sort of objection by insisting that we have not given reason to think that the use of machine learning techniques must *in principle* raise the specter of moral tradeoffs. With sufficiently large sets of training data and sufficient processing power, engineers could ensure whatever desired behavior in every instance through the use of machine learning. This is intuitive. Return again to our dog trainer: if she had all the time she needed and a sufficiently intelligent dog, she could ensure that it was behaving exactly as she

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

wanted in any scenario merely through the use of operant conditioning. The difficulty for both the engineer and the dog trainer is that *sufficient* training data and processing power to *guarantee* outcomes in this way is far beyond their reach. Indeed, it appears to us that it would require infinite or near-infinite training time, data, and power with which to process it. In the case of machine learning, it appears that only a training regime with those characteristics would allow an engineer to *ensure* that the input-output mappings that a machine learner generates are exactly the same as they would be if they were explicitly coded. Using our dog trainer to illustrate one final time, consider just *how much* operant conditioning it would take to guarantee the behavior of even an incredibly intelligent dog for any scenario that might arise.

We are a long way from having sufficient training data and processing power for this observation about the theoretical possibilities of machine learning to be a useful response to our objection, but it does highlight how one might appeal to more realistic technological possibilities to undermine our case for pessimism: to the extent that we can effectively divorce how AVs will behave in accident scenarios that are structurally like trolley cases from training for other behaviors, we can effectively treat accident scenarios as if they are being programmed according to the traditional paradigm.[17] The objector can concede that Regime is an ethical question that deserves our attention, but insist that if we can address the behavior of AVs in accident scenarios independently of how they otherwise behave, we would be acting wrongly by not thinking explicitly about how to train for or program specific behaviors in particular accident scenarios.

Notice that an objection of this form undermines our argument only if and to the extent that there is some reason to think that there are technological options available that actually allow developers to treat the accident scenarios that are structured like trolley cases in a way that is totally distinct from both other accident scenarios and typical driving scenarios. One reason to be optimistic that there are or will be such options is that there are many tools in an engineer's toolkit for trying to solve the problem: they can choose from a variety of different types of machine learners (e.g., Bayesian vs. evolutionary learners), they can vary the training approach (e.g., using reinforcement vs. supervised learning), they can vary the composition of the training set, they can manually tweak the algorithms to make adjustments to driving behavior, and so on.

The approach that provides the best hope to reject our argument on technological grounds is to try to "divide and conquer." Many problems that machine learning is deployed to solve are best managed by dividing the problem into subproblems and then creating individual training regimes to solve those independently. In the case of AVs, for example, a better driving system might be produced by training algorithms for managing highway and city driving separately than would be if we attempted to train a single driving algorithm. Similarly, it

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

might be possible to deploy the divide-and-conquer approach with respect to accident scenarios and ordinary driving scenarios. One could imagine trying to train a system only for managing behavior in accident scenarios that are like trolley cases and then working to ensure that the AV deploys that subsystem only when appropriate. To the extent that the system can reliably detect when it is in the relevant sort of situations, developers will have safeguarded against the concern that by trying to address accident scenarios they may have compromised the overall reliability of AVs.

We are skeptical that the divide and conquer approach threatens our position. For the approach to work, developers not only need to be able to develop training sets or write traditional algorithms that generate reliable behavior in the relevant accident scenarios, but they also need to ensure that the system can reliably detect when it is in those accident scenarios. These are serious technical challenges. Unlike traditional driving scenarios for which data can be collected by monitoring actual drivers, we do not have an obvious source of training data for accident scenarios. Training data for accident scenarios that are structured like trolley cases will have to be manufactured or simulated, raising questions about whether training using simulated data will actually yield reliable results. Furthermore, there is the technical challenge of generating a sufficient amount of training data. As we noted above, we do not want the algorithm to, for example, save the five instead of the one, but instead to save the many over the few. More importantly, integrating the algorithms for accident scenarios and other driving scenarios also involves a training regime; we must train an algorithm to detect or make a decision about which kind of scenario it is in, and we again have to make a choice about how to structure the training regime when there are potential tradeoffs.[18] For this reason, we believe that the divide-and-conquer approach could solve the problems we identified for Trolley Optimism only to reintroduce them elsewhere in the design process.

Further, even if there is a technical solution that allows us to manage for the relevant accident scenarios separately, this is not sufficient to justify the answer to Behavior that the Trolley Optimist takes to be correct. Our objection to Premise 2 of The Optimist's Argument highlights the importance of Regime. These technological solutions, at best, make it possible that we *can* train for specific behaviors in specific scenarios but not that we *should*. If we should, as we argued above, it is because our best ethical theorizing about how to answer Regime tells us that we should and not because there is a structural similarity between some accident scenarios and trolley cases. Again, what is crucial for substantiating Trolley Pessimism in the way that we are attempting to do so is the fact that the range of engineering options open to Annabelle at the design stage makes it possible that there are a set of moral considerations that bear on what she is permitted to do

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

that are not directly related to the considerations triggered by the structure of potential accident scenarios.

Philosophical Objections

In advancing our case against Trolley Optimism, we set as our target the second premise of The Optimist's Argument.[19] However, a possible objection to our view has it that we are actually mistaken about how the machine learning considerations we have adduced bear on The Optimist's Argument: whereas we take ourselves to be undermining the second premise, we have actually given reason for rejecting the first.

The reasoning behind this objection is easy to understand. We said above that in deploying machine learning techniques to govern the driving of AVs in accident scenarios, engineers risk tradeoffs involving AV driving performance outside of accident scenarios. This suggests that there are moral considerations that bear on Annabelle's case that are not present in trolley cases, which suggests that accident scenarios are not structurally similar to trolley cases. If so, then first premise is false.

In response to this objection, we want to be both somewhat stubborn and somewhat concessive. Strictly speaking, we do not agree that the presence of moral considerations that bear on Annabelle's case that do not bear on trolley cases shows that accident scenarios cannot be structurally similar to trolley cases. In explaining structural similarity above, we were careful to characterize it in terms of the *number* of routes available, the *way* in which those routes are traversable, and the moral implications for *which route is traveled*. The argument that we have advanced against Trolley Optimism focuses not on a difference among one of these dimensions between trolley cases and accident scenarios, but instead on a difference about *what it takes to cause an AV to travel one of the routes*. Organizing things in this way, we think, puts us in the best position to appreciate what has so far gone unnoticed in the literature: that the engineering choices that will ultimately give rise to AV driving behavior must themselves be scrutinized. Further, it also does justice to the way that Trolley Optimists appear to reason, beginning as they do with an observation about the circumstances of possible collisions involving AVs and the similarity between *those circumstances* and various trolley cases.

We are content to concede, though, that this organizational choice may seem to some readers as just that. Unsurprisingly, one must point to *some* difference between Annabelle's context and a trolley case to show that reflection on trolleys cannot directly settle the question that confronts her. If one wants to categorize putative differences so that they are all of a kind, we have no principled objection

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

to doing so (though, again, we do not concede that this is the most natural or helpful way of proceeding). In that case, we would nevertheless want to insist that the *technological* difference between how outcomes are realized in traditional algorithmic systems and machine learning systems is of paramount importance because of the ethical dimensions of AV design, and indeed the design of any system dependent on machine learning, that it reveals.

A final philosophical objection remains to be addressed. The Trolley Optimist could insist that generating a specific behavior in accident scenarios that mirror trolley cases is of *overriding* moral importance. According to this response, Behavior is the fundamental question of AV design, it should be answered in the affirmative, and we should ensure the relevant behaviors come what may—even to the point of abstaining from deploying AVs at all if their driving cannot be guaranteed to conform with the appropriate verdicts about trolley cases.

This response transforms Trolley Optimism into a form of *Trolley Authoritarianism*. We take the authoritarian view to be not only implausible on its face but also disconnected from what motivates actual Trolley Optimists. Whatever plausibility Trolley Authoritarianism may appear to have is likely the result of a misconception that we have been at pains to illuminate. Even if trolley cases could be used to show that there exist hard constraints on the ways in which it is permissible to realize certain valuable outcomes, those constraints apply to the design and deployment of AVs *only if they bear on the programming options that are available to engineers like Annabelle*, and that cannot be substantiated by showing that accident scenarios resemble trolley cases.

Implications and Conclusions

We took as our focal point in this paper Annabelle and her moral deliberation about how to program an AV. We asked whether reflection on trolley cases would be especially useful to her, given the similarities between some of those cases and potential accident scenarios involving AVs. Arguing that those cases are not especially relevant required us to attend to some details about how machine learning techniques are used in designing autonomous cars. The reader may be surprised to learn, in light of all of that, that we do not think the most important lesson of our argument has anything essentially to do with either cars or trolleys. The lesson is instead a perfectly generalizable one about how to do applied ethics where machine learning is involved.

The mistake of the Trolley Optimist is focusing exclusively on a moral evaluation of the *behavior* of a piece of technology, rather than on the *actions of people* that cause the behavior. And that mistake is an easy one to make as technology becomes more complex, and in particular as machine learning is deployed across

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

an ever-growing range of domains. That is because it will become increasingly common for technology to behave in ways that mirror possible actions that a person might take, inclining one to take the behavior as a fitting or even the primary object of moral evaluation (as we would with a person's action). We have attempted to show what precisely is mistaken about this, using the case of AVs. Even if we are badly mistaken, though, about the ways in which machine learning techniques are used in the case of AVs, we think that the general lesson stands. As theorizing about the ethics of automated technology moves forward, we must be on guard to avoid faulty reasoning that moves us from an evaluation of technological behavior to a conclusion about the permissibility of engineering actions.[20] Doing so will require ethicists to pay close attention to the particular options available to engineers like Annabelle.[21]

Notes

1. For a selection of popular media pieces, see Achenbach (2015), Doctorow (2015), Donde (2017), Hao (2018), Lin (2013), Marshall (2018), and Worstall (2014). In the academic literature, see Bonnefon et al. (2019), Lin (2015), and Wallach and Allen (2009).
2. Himmelreich (2018).
3. For a sampling of arguments in the philosophical literature that do concern policy, see the following: Gogoll and Müller (2017), Hevelke and Nida-Rümelin (2015), Himmelreich (2018), and Mladenovic and McPherson (2016).
4. We are thinking especially of the US legal context, in which there are no federal regulations squarely governing AVs. The Safely Ensuring Lives Future Deployment and Research in Vehicle Evolution Act (SELF DRIVE Act) may become the first, having passed the House in a 54–0 vote on September 6, 2017. See H.R.3388—SELF DRIVE Act, available at https://www.congress.gov/bill/115th-congress/house-bill/3388. The SELF DRIVE Act does not speak with specificity to how to program AVs. Instead, it mainly memorializes the state-preempting, federal role in ensuring AV safety as it relates to "design, construction, and performance, by encouraging the testing and deployment of such vehicles." See §2 (Purpose) and §3 (NHTSA Authority and State Preemption for Autonomous Motor Vehicles). Illustrating the lack of specific directives, §4(a) requires that "[n]ot later than 24 months after the date of the enactment of this section, the Secretary of Transportation shall issue a final rule requiring the submission of safety assessment certifications regarding how safety is being addressed by each entity developing a highly automated vehicle or an automated driving system." This language does not specify how safety is to be addressed, but merely requires the Secretary to compel developers to reveal how safety is in fact being addressed.
5. Those we take to be Trolley Optimists (see, for example, Lin [2013]; Lin [2015] Noothigattu et al. [2017]) do not themselves explicitly characterize their views this

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

way. Our characterization is designed to accommodate a variety of forms of the view found in different literatures and in the popular media. However, a commitment to this inference is implicit in the directives issued and lessons drawn from the use of trolley cases. For a similar characterization of such views, see Hübner and White (2018).

6. We should note that it may take a bit of imaginative work to develop even putatively structurally similar accident scenarios and trolley cases so long as whichever case with which we begin is sufficiently complex. To illustrate that it is not only simple cases of the sort discussed in the main text that could be offered in support of Premise 1, we provide an additional example just below. We leave it to the reader to consider whether there may be some trolley cases that do not allow for even apparently parallel accident scenarios. We are in agreement with Frances Kamm (2020) that some discussions of accident scenarios present them as if they are structurally similar to trolley cases, when in fact they are not.

   Probably the most famous trolley case after the standard case is that of the Large Man. In Large Man, a trolley is again on course to strike and kill five innocent people on the track. As in the standard case, a bystander can prevent their deaths, but in this scenario only by pushing a very large man onto the track; this will kill the large man, but the collision will slow the trolley enough to prevent it from reaching the five. What kind of accident scenario might resemble this situation? What is needed is a case in which an AV can avoid hitting five innocents only by colliding into another that is moved into its path. It may be hard to imagine that latter feature being present in a case in which the autonomous technology is itself manipulating the environment and selecting one of the available options. Perhaps it might happen like this: An unoccupied vehicle is traveling north toward a four-way intersection. Five pedestrians unexpectedly enter the crosswalk beyond the intersection, and there is no opportunity for the car to swerve or to slow in time. A second AV with one passenger onboard is approaching the same intersection from the west. Because the vehicles are operated via a shared network, it is possible for the systems that govern them to coordinate their driving. In this case, the combined mass of the second car and its occupant is just enough to halt the momentum of the first if they collide, thus preventing the five from being hit. All that the system need do is accelerate the second car into the intersection before the first passes through. For related technical material, see Dresner and Stone (2008).
7. See Noothigattu et al. (2017). For a slightly different version of the Democratic variant that also relies on aggregating people's verdicts on trolley cases as part of setting policy for how AVs should behave, see Savulescu et al. (2021).
8. For a discussion of some other approaches to defending Trolley Pessimism along with some challenges to those approaches, see the Appendix in this volume.
9. This is not to say that programmers of traditional algorithms can predict the output that corresponds to every input or that the outputs are deterministic. What is characteristic of the traditional paradigm is that programmers specify precisely the steps for generating the output.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

10. This description provides only a rough model of how machine learning works. It roughly captures how reinforcement and supervised learning work, but it omits various details. For example, a complete explanation of machine learning would include discussion of the *test set*, a set of data on which machine learners are evaluated after having learned by way of the training set. In light of the evaluation, the learner's outputted algorithm can then be modified manually or through additional training. Furthermore, the description in the main text does not capture unsupervised learning, a method by which a machine learner is used to find correlations in some large data set. Supervised, unsupervised, and reinforcement learning are the three most significant kinds of machine learning in use today (though, in application, these categories are often not perfectly sharp and the learning methods can be used in conjunction with one another).

    In general, we have approached this section under the assumption that many readers are not closely familiar with the techniques used in machine learning applications, and so have attempted to emphasize clarity with respect to the features of machine learning that are most relevant for our argumentative aim. For readers interested in an introduction to the technical foundations of machine learning, we recommend Russell and Norvig (2015, section V). For a less technical, but still technically grounded and accurate introduction, we recommend Broussard (2018, chapter 7).
11. For ease of exposition, we have sometimes written in this section as if a single algorithm determines the driving behavior of an AV, which need not (and almost certainly would not) be the case. We will consider the significance of this in the fourth section, and again revisit in the final section the implications of our work if we are mistaken about some of the details involving the design of AVs.
12. Some of our reasoning in this section may have been anticipated by Himmelreich (2018). His remarks on the topic are somewhat underdeveloped, though, as they play a relatively minor role in his argumentative project; this makes it difficult to tell to what degree we are in agreement. At the very least, we applaud and concur with Himmelreich's observation that whether trolley cases are directly relevant to programming AVs will depend in part on how AVs are engineered.
13. Lin (2013).
14. It might be tempting to think that adding additional cases surely could do nothing other than increase performance, but that thought is mistaken. One reason is that providing more and more cases could lead to overfitting the data, causing worse performance in cases outside the training set. More importantly, the choice to design a training regime oriented around incorporating accident scenarios is not simply a choice of including additional data but also a choice about what programmers want a particular model to accomplish and what classes of data are relevant to advancing that end. In the case of AVs, machine learning is primarily used for object detection and recognition. The detection tasks required for everyday driving scenarios are radically different from those required for detection in trolley-like accident scenarios; whereas for typical driving, detecting how many people are in a parked vehicle is

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

irrelevant, for an accident scenario in which an AV must hit one of two parked cars, it may be extremely relevant. Therefore, there is a potential cost to concentrating on accident scenarios when training an object-detection algorithm, because the goals of object detection may vary between accident scenarios and ordinary driving conditions. Furthermore, issues can arise when there are so-called class imbalances in the data, where there are more instances of one class of data than another. In such cases, algorithms are more likely to ignore cases in the classes with fewer instances (see Johnson and Khoshgoftaar [2019]). This has implications for Regime in that it might be that in order to create algorithms responsive to accident scenarios, the proportion of training resources dedicated to them might need to be significant, with a potentially significant impact on algorithmic performance in nonaccident scenarios.

15. Or the distinction between killing and letting die, or the distinction between intended and merely foreseen consequences, or the significance of using a person as a mere means, or whatever considerations one takes to make a moral difference between cases like the traditional trolley case and Large Man. We intend for the example in the main text to be merely illustrative.
16. Kamm (2020) agrees that such explanations matter for the ethics of AVs, as when she writes, "It is interesting and important that which theory correctly grounds nonconsequentialist prohibitions on harming could be relevant to which principles to use in programming machines" (90). However, she seemingly denies that this is because the explanations for the deontic verdicts in trolley cases might themselves apply to people in situations like Annabelle's and thinks instead that it is because the explanations might apply directly to the behavior of vehicles. This is suggested when she writes that her "concern here is not *how* to make self-driving cars behave properly but *what* the proper way for them to behave is" (88, emphasis in original), and when she extends that thought further two pages later, writing, "We could at least say that there is a reason for there not to be entities that would kill someone in this [a morally impermissible] way and people who were aware of these reasons might be morally required to interfere with the machine that would kill in this way. This could be true even if the machine was not designed by people but fell like manna from heaven or grew like a plant" (90).
17. We hope that the reader does not doubt that engineers have nothing resembling infinite training data and computational power at their disposal. Those who remain optimistic that engineers may nevertheless have something approaching enough of both might consider the myriad ways in which existing AVs fail in even mundane driving scenarios. See, for example, Broussard (2018, ch. 8).
18. This last consideration also undermines attempts to solve the problem by reverting to the traditional algorithm paradigm. While it might be easier to explicitly program rules for accident scenarios than to develop a sufficiently robust training regime to ensure certain behaviors, there remains the problem of being able reliably to deploy traditional algorithms in just the right conditions.
19. In the Appendix, we bolster the case for thinking that this is the most promising target, casting a skeptical eye over the charges raised against Premise 1.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

20. For this reason, we may be correct that The Optimist's Argument fails even if we have reasoned from mistaken assumptions about AVs and machine learning. Whether or not the technological case we are advancing against The Optimist's Argument will apply when machine learning is not heavily implicated in the behavior of AVs will depend again on whatever engineering options are genuinely open. If there are reasons that make it a nontrivial engineering task to guarantee conformity with verdicts about trolley cases, then it is possible that an argument parallel to the machine learning argument that we offer here will apply. Indeed, we are inclined to think that this is most likely the case. When building mathematical models of the sort that could govern the executive functioning of complex systems like AVs, engineers who choose not to use machine learning at this stage must make a series of decisions about the design of the models and the processes for testing them. As with the engineering options that we highlight in the main text, these options too appear to be subject to moral evaluation. We are especially grateful to Chris Beck for urging us to consider this issue, and for helping us to think it through.
21. We gratefully acknowledge the contributions of Mark Lee, who assisted us with research during the earliest stages of this project. For providing comments on earlier versions of the paper, we thank Selim Berker, Barbara J. Grosz, Johannes Himmelreich, Duncan Purves, Gina Schouten, Gili Vidan, and the editors of this volume. Sven Nyholm provided helpful comments late in the production cycle of this volume, and we regret that time and space considerations did not permit us to address all of the valuable insights he shared. We have been fortunate to present this work at a variety of venues. We are grateful for helpful feedback from audiences at the 2018 Rocky Mountain Ethics Congress; Harvard University; Tufts University; Northeastern University; the 2019 International Conference on Technology, Knowledge, and Society; the 2019 US Frontiers of Engineering Symposium; and the University of Toronto Centre for Ethics. Some of Basl's time working on this paper was funded by a grant for work on human-robot collaboration funded by the National Science Foundation (NSF-1928654). Behrends was supported by a fellowship from the Edmond J. Safra Center for Ethics at some of the time during which he worked on this paper.

References

Achenbach, Joel. 2015. "Driverless Cars Are Colliding with the Creepy Trolley Problem." *The Washington Post*, December 29, 2015. https://www.washingtonpost.com/news/innovations/wp/2015/12/29/will-self-driving-cars-ever-solve-the-famous-and-creepy-trolley-problem/?utm_term=.8c16a3ee1a28.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2019. "The Trolley, the Bull Bar, and Why Engineers Should Care About the Ethics of Autonomous Cars." *Proceedings of the IEEE* 107, no. 3: 502–4.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Understand the World*. Cambridge, MA: The MIT Press.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

Doctorow, Cory. 2015. "The Problem with Self-Driving Cars: Who Controls the Code?" *The Guardian*, December 23, 2015. https://www.theguardian.com/technology/2015/dec/23/the-problem-with-self-driving-cars-who-controls-the-code

Donde, Jay. 2017. "Self-Driving Cars Will Kill People. Who Decides Who Dies?" *Wired*, September 21, 2017. https://www.wired.com/story/self-driving-cars-will-kill-people-who-decides-who-dies/.

Dresner, Kurt, and Peter Stone. 2008. "A Multiagent Approach to Automated Intersection Management." *Journal of Artificial Intelligence Research* 31: 591–656.

Gogoll, Jan, and Julian F. Müller. 2017. "Autonomous Cars: In Favor of a Mandatory Ethics Setting." *Science and Engineering Ethics* 23, no. 3: 681–700.

Hao, Karen. 2018. "Should a Self-Driving Car Kill the Baby or the Grandma? Depends Where You're From." *MIT Technology Review*, October 24, 2018. https://www.technologyreview.com/s/612341/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/.

Hevelke, Alexander, and Julian Nida-Rümelin. 2015. "Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis." *Science and Engineering Ethics* 21, no. 3: 619–30.

Himmelreich, Johannes. 2018. "Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations." *Ethical Theory and Moral Practice* 21, no. 3: 669–84.

Hübner, Dietmar, and Lucie White (2018). "Crash Algorithms for Autonomous Cars: How the Trolley Problem Can Move Us Beyond Harm Minimisation." *Ethical Theory and Moral Practice* 21, no. 3: 685–98.

Johnson, J. M., and T. T. M. Khoshgoftaar. 2019. "Survey of Deep Learning with Class Imbalance." *Journal of Big Data* 6: 27. https://link.springer.com/content/pdf/10.1186%2Fs40537-019-0192-5.pdf.

Kamm, Frances. 2020. "The Use and Abuse of the Trolley Problem: Self-Driving Cars, Medical Treatments, and the Distribution of Harm." In *Ethics of Artificial Intelligence*, edited by Matthew Liao, 79–108. Oxford: Oxford University Press.

Lin, Patrick. 2013. "The Ethics of Autonomous Cars." *The Atlantic*, October 8, 2013. https://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lin, Patrick. 2015. "Why Ethics Matters for Autonomous Cars." In *Autonomous Fahren: Technische, rechtlische und gesellschaftliche Aspekte*, edited by Markus Maurer, J. Christian Gerdes, Barbara Lenz, and Herman Winner, 69–86. Berlin: Springer.

Marshall, Aarian. 2018. "What Can the Trolley Problem Teach Self-Driving Car Engineers?" *Wired*, October 24, 2018. https://www.wired.com/story/trolley-problem-teach-self-driving-car-engineers/.

Mladenovic, Milos N., and Tristram McPherson. 2016. Engineering Social Justice into Traffic Control for Self-Driving Vehicles? *Science and Engineering Ethics* 22, no. 4: 1131–49.

Noothigattu, Ritesh, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. 2018. "A Voting-Based System For Ethical Decision Making." In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, no. 1.

Russell, Stuart J., and Peter Norvig. 2015. *Artificial Intelligence: A Modern Approach* (3rd ed.). Noida: Pearson India Education Services.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026

Savulescu, Julian, Christopher Gyngell, and Guy Kahane. 2021. "Collective Reflective Equilibrium in Practice (CREP) and Controversial Novel Technologies." *Bioethics* 35, no. 7: 652–63. doi:10.1111/bioe.12869.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Worstall, Tim. 2014. "When Should Your Driverless Car from Google Be Allowed to Kill You?" *Forbes*, June 18, 2014. https://www.forbes.com/sites/timworstall/2014/06/18/when-should-your-driverless-car-from-google-be-allowed-to-kill-you/#2e21fba6fa5b.

Downloaded from academic.oup.com/book/44058/chapter/371949049 by Hebrew University of Jerusalem user on 10 September 2026