1 AUTONOMOUS VEHICLES AND MORAL UNCERTAINTY

Vikram Bhargava and Tae Wan Kim

Autonomous vehicles have escaped the confines of our imaginations and found their way onto our roadways. Major companies, including GM, Nissan, Mercedes-Benz, Toyota, Lexus, Hyundai, Tesla, Uber, and Volkswagen, are developing autonomous vehicles. Tech giant Google has reported that the development of autonomous vehicles is among its five most important business projects (Urmson 2014). Autonomous cars are here, not going away, and are likely to become increasingly prevalent (Fagnant and Kockelman 2013; Goodall 2014a, b).

As the number of autonomous vehicles on roadways increases, several distinctively philosophical questions arise (Lin 2015):

> *Crash*: Suppose a large autonomous vehicle is going to crash (perhaps due to hitting a patch of ice) and that it is on its way to hitting a minivan with five passengers head on. If it hits the minivan head on, it will kill all five passengers. However, the autonomous vehicle recognizes that since it is approaching an intersection, on the way to colliding with the minivan it can swerve in such a way that it first collides into a small roadster, thus lessening the impact on the minivan. This would spare the minivan's five passengers, but it would unfortunately kill the one person in the roadster. Should the autonomous vehicle be programmed to first crash into the roadster?

This scenario of course closely resembles the famous trolley problem (Foot 1967; Thomson 1976).[1] It also raises a question at the intersection of moral philosophy, law, and public policy that is unique to autonomous vehicles. The question is, who should be able to choose the ethics for the autonomous vehicle—drivers, consumers,

passengers, manufacturers, programmers, or politicians (Lin 2015; Millar 2014)?[2] There is another question that arises even once we settle who ought to be able to choose the ethics for the autonomous vehicles:

> *The Problem of Moral Uncertainty*: How should autonomous vehicles be programmed to act when the person who has the authority to choose the ethics of the autonomous vehicle is under moral uncertainty?

Roughly, an agent is morally uncertain when she has access to all (or most) of the relevant non-moral facts, including but not limited to empirical and legal facts, but still remains uncertain about what morality requires of her. This chapter is about how the person who is ultimately chosen to make the decisions in the *Crash* scenario can make appropriate decisions when in the grip of moral uncertainty. For simplicity's sake, in this chapter we assume this person is a programmer.

Decisions are often made in the face of uncertainty. Indeed, there is a vast literature on rational decision-making under uncertainty (De Groot 2004; Raiffa 1997). However, this literature focuses largely on empirical uncertainty. Moral uncertainty, on the other hand, has received vastly less scholarly attention. With advances in autonomous vehicles, addressing the problem of moral uncertainty has new urgency. Our chief purpose in this chapter is a modest one: to explore the problem of moral uncertainty as it pertains to autonomous vehicles and to outline possible solutions to the problem. In section 1.1, we argue that the problem is a significant one and make some preliminary remarks. In section 1.2, we critically engage with two proposals that offer a solution to the problem of moral uncertainty. In section 1.3, we discuss a solution that we think is more promising, the solution provided by the philosopher Andrew Sepielli. In section 1.4, we offer some support in its defense. We conclude in section 1.5.

1.1 Motivation and Preliminaries

Let's return to the *Crash* scenario. Suppose Tegan, a programmer tasked with deciding the appropriate course of action in the *Crash* scenario, thinks she should program the autonomous vehicle to collide into the roadster on the way to the minivan, under the consequentialist rationale that the roadster has fewer passengers than the minivan. She hesitates because she recalls her ethics professor's deontological claim that doing so would be seriously wrong[3]—it would use the one passenger in the roadster as a mere-means in a way that is morally impermissible. Tegan is not persuaded by her professor's prior guidance and gives 90% credence (subjective probability) to the view that she should program the vehicle to first crash into the roadster; she gives only 10% credence to her professor's conclusion that she should not crash into the roadster on the way to the minivan.[4]

From Tegan's own perspective, how should she program an autonomous vehicle to deal with *Crash*-like scenarios? Tegan faces the problem of moral uncertainty.

Caveat: Any real driving situation an autonomous vehicle will face is likely to be much more complicated than the foregoing scenario. Nevertheless, the scenario contains enough relevant factors for our purposes. For Tegan, moral arguments are the source of normative uncertainty, but it is worth noting that other types of normative views (e.g., legal, cultural, religious) can play similar roles, creating prescriptive uncertainty. Also, Tegan faces just two competing arguments, while many decision makers can face more than two. For simplicity's sake, however, we discuss scenarios in which two arguments are competing, although some of the frameworks we discuss can, in principle, accommodate scenarios with more than two competing arguments.

The problem of moral uncertainty derives primarily from the following conditions: (1) the two normative propositions corresponding to Tegan's decision—"I should program the vehicle to crash into the roadster" and "I should not program the vehicle to crash into the roadster"—are mutually exclusive, and (2) Tegan's credence is divided between the two propositions—she is uncertain about which of the two propositions is true (Sepielli 2009).[5] Put another way: even if Tegan is certain about a range of empirical facts, she may still remain uncertain about the reasons that those very facts give her with respect to what to do (Sepielli 2009).

We acknowledge that a solution to *Crash* is not complete unless we offer a plausible framework of decision-making under empirical uncertainty (e.g., Hare 2012). We assume for now that the solution we discuss can coherently be combined with the best account of decision-making under empirical uncertainty—ascertaining whether this assumption is in fact true is a promising future avenue of research.[6]

Tegan requires a meta-normative framework to adjudicate between the normative prescriptions of competing moral theories. In this chapter, we argue for the importance of such a framework and encourage scholars of robot ethics to pay more attention to the problem of moral uncertainty. But first, it is worth dealing with a thought that might be lurking in some readers' minds, namely, that the very notion of moral uncertainty is a misguided one. Specifically, some readers might be thinking that of the two competing moral arguments, only one of them is right. Therefore, there is no uncertainty *ab initio* and the problem is only apparent. For instance, if it is in fact morally true that one should never use another as a mere-means, then Tegan should not program the car to first crash into the roadster.

We agree that there may be no moral uncertainty from the perspective of objective reason or objective "should" (Harman 2015). Moreover, we do not deny the importance of figuring out the objectively correct answer, assuming

one exists—that is, that ethics is not relative. But it is important to note that the aforementioned concern is consistent with the view we defend. Though programmers should strive to ascertain the objectively correct answer, this does not eliminate the fact that a decision might have to be made prior to one's having secured the objectively correct answer. Tegan, in the above scenario, cannot make her decision purely on the basis of objective reasons, since her doxastic state is already plagued with moral uncertainty. Yet she needs to decide how to program the car. The given reality is that Tegan cannot help but base her decision on her degree of belief in a moral view—that is, from her representation of the objective "should" (Sepielli 2009). Thus Tegan ought to make the best decision given *her* degree of belief in the relevant normative prescription. So Tegan requires an additional decision framework, one that is not designed primarily for objective reason or objective "should"—that is, a framework that begins with her own uncertain normative beliefs but still helps her make more appropriate and rational decisions.

1.2 Two Possibilities

We now consider (and ultimately reject) two proposals for making decisions under moral uncertainty. The first proposal—the "Continue Deliberating" view—suggests that Tegan should not make a decision; instead, she should continue deliberating until she figures out what morality requires of her. We are sympathetic to this position. Indeed, we too think that programmers should continue to deliberate about moral problems insofar as they are able. Nevertheless, we believe that there are circumstances in which programmers may lack the luxury of time or resources to continue deliberating but must nevertheless decide how to act. Tegan might deliberate for some time, but she cannot put all of her time and effort into figuring out what to do in *Crash* and will need to make a decision soon enough.

Perhaps more important, continuing to deliberate is in some contexts, in effect, making a decision about one of the very choices the programmer is uncertain about. For example, if Tegan opts not to program the autonomous vehicle to first swerve into the roadster, she in effect already commits to the prescription of one of the moral views. That is, if she decides not to program the autonomous car to first swerve into the roadster, she rejects the prescription of the consequentialist view and allows more lives to be killed. Inaction is often a choice, and it is typically a choice of status quo. The "Continue Deliberating" view lacks the resources to explain why the existing state of affairs is the most appropriate choice.

The second proposal—call it the "My Favorite Theory" view—is an initially tempting response to moral uncertainty (Gustafsson and Torpman 2014). That

is, do what the conclusion of the normative argument you think is most likely to be correct tells you to do. For instance, if Tegan thinks the consequentialist prescription to crash into the roadster is most likely true, then she should program the car to do just that. But this view has some problems, an analysis of which will yield an important condition any adequate solution to the problem of moral uncertainty must meet. We can better understand this problem by considering a parallel problem in empirical uncertainty (Sepielli 2009). Consider a hypothetical variant of the real case of the Ford Pinto:

> *Pinto*: The CEO of Ford is deciding whether to authorize the sale of its recently designed hatchback car, the Pinto. She is not sure how to act, because she is empirically uncertain about how the Pinto's fuel tank will affect the life of its drivers and passengers. After reading a crash-test report, she has a 0.2 degree of belief that the Pinto's fuel tank will rupture, causing a potentially fatal fire if the car is rear-ended, and a 0.8 degree of belief that there will not be any such problems. Thinking she should go with what she thinks is most likely—that there will not be any problems with the fuel tank—she authorizes the sale of the Pinto.

Here the CEO clearly makes a poor choice. One cannot simply compare 0.2 and 0.8. One must consider the value of the outcomes. Of course, car designs cannot be perfect, but a 20% probability of a life-threatening malfunction is obviously too high. The CEO failed to weigh the consequences of the actions by their respective probabilities. If she had taken into consideration the value of the outcomes, it would not have made sense to authorize the sale of the Pinto. A similar problem applies in the moral domain—the weight of the moral value at stake must be taken into consideration.

For instance, returning to the situation Tegan faces, even if she thinks that the proposition "I should program the vehicle to crash into the roadster" is most likely true, it would be a very serious wrong if the competing proposition, "I should not program the vehicle to crash into the roadster," is correct, since treating someone as a mere-means would be a serious deontic wrong. In other words, though Tegan thinks that her professor's view is mistaken, she recognizes that *if* her professor's arguments are correct and she nevertheless programs the car to first crash into the roadster, then she would commit a serious deontological wrong.

As such, an adequate solution to the problem of moral uncertainty must take into account the moral values associated with the particular normative proposition, weighted by their respective probabilities, not merely the probability that the normative proposition in question is true. Another way to put the point is that a programmer, in the face of moral uncertainty, must hedge against the view

with the greater moral value at stake or meet what we shall call the "expected *moral* value condition," which we apply to programmers in the following way:

> *The Expected Moral Value Condition*: Any adequate solution to the problem of programming under moral uncertainty must offer the resources by which a programmer can weigh the degree of moral harm, benefit, wrong, right, good, bad, etc., by their relevant probabilities.[7]

On an account of moral uncertainty that meets this condition, there may very well be instances when a programmer should act according to what she considers the less probable normative view because the less probable normative view has something of significant moral value at stake. But we've gotten ahead of ourselves. Making this sort of ascription requires being able to compare moral value across different moral views or theories. And it is not obvious how this can be meaningfully done. In the next section, we will elucidate what we think is a promising approach for making comparisons of moral value across different moral views.

1.3 An Expected Moral Value Approach

Suppose Tegan is convinced of the importance of weighing the moral value at stake and decides she wants to use an expected moral value approach to do so.[8] In other words, Tegan must figure out the expected moral value of the two mutually exclusive actions and choose the option that has the higher expected moral value. But she soon realizes that she faces a serious problem.

Tegan might have some sense of how significant a consequentialist good it is to save the five lives in the minivan, and she also might have some sense of how significant a deontological bad it is to use another as a mere-means (namely, the person in the roadster); but still, troublingly, she may not know how the two compare. It is not clear that the magnitude of the moral value on the consequentialist view is commensurable with the magnitude of the moral value on the deontological view. This has been called the "Problem of Inter-theoretic Value Comparisons" (PIVC) (Lockhart 2000; Sepielli 2006, 2009, 2013; MacAskill, forthcoming).

The PIVC posits that moral hedging requires comparing moral values across different normative views. And it is not obvious that this can be done. For example, it is not clear how Tegan can compare the consequentialist value of maximizing net lives saved with the deontic wrong of using someone as a mere-means. Neither consequentialist views nor deontological views themselves indicate how to make inter-theoretic comparisons. Any adequate expected value proposal must explain how it will handle this problem.[9]

Although the PIVC is thorny, we maintain it can be overcome. We find Sepielli's works (2006, 2009, 2010, 2012, 2013) helpful for this purpose. Significantly, Sepielli points out that just because two things cannot be compared under one set of descriptions does not mean they cannot be compared under an analogous set of re-descriptions. Sepielli's nuanced body of work is more complex than the following three steps that we articulate. Still, these three steps form the essence of what he calls the "background ranking approach" and are also helpful for our investigation.[10] Using the background ranking approach to solve the problem of inter-theoretic value comparisons is simple at its core, but its execution may require practice, much as with employing other rational decision-making tools.

Step 1: The first step involves thinking of two morally analogous actions to programming the vehicle to crash into the roadster. The first analogy should be such that, if the moral analogy were true, it would mean that crashing into the roadster is better than not crashing into the roadster. The second analogy should be such that, if the moral analogy were true, it would mean that not crashing into the roadster is better than crashing into the roadster. Suppose the first analogy to crashing into the roadster is donating to an effective charity that would maximize lives saved, instead of donating to a much less effective charity that would save many fewer lives. (In this case, the analogy is in line with the consequentialist prescription. It is a decision strictly about maximizing net lives saved.) Call this the "charity analogy." Suppose the second analogy to crashing into the roadster is a doctor killing a healthy patient so she could extract the patient's organs and distribute them to five other patients in vital need of organs. (In this case, the analogy is in line with the deontological prescription of not using a person as a mere-means.) Call this the "organ extraction analogy." Note that performing this step may require some moral imagination (Werhane 1999), skill in analogical reasoning, and perhaps even some familiarity with what the moral literature says on an issue.

Step 2: Identify one's credence in the two following mutually exclusive propositions: "I should program the vehicle to crash into the roadster" and "I should not program the vehicle to crash into the roadster." As stated earlier, Tegan has a 0.9 credence in the first proposition and a 0.1 credence in the second proposition.

Step 3: The third step involves identifying the *relative* differences in the magnitude of the moral value between the two propositions from *Step 2* on the assumption that each of the analogies from *Step 1* in question holds. Let's call the difference in the moral value of programming the vehicle to crash into the roadster versus not doing so, given the charity analogy is true, "W." Suppose then that the difference in moral value of programming the vehicle to crash into the roadster versus not doing so, given the organ extraction analogy is true, is "50W" (i.e., the difference in moral value is fifty times that of the difference in moral value associated with the charity analogy).

Keep in mind that Tegan can do this because her views about the relative differences in the magnitude of the moral value between the two propositions, conditional on each analogy holding true, can be *independent* of her beliefs about the two mutually exclusive prescriptions of "I should program the vehicle to crash into the roadster" and "I should not program the vehicle to crash into the roadster." As Sepielli notes, "Uncertainty about the ranking of a set of actions under one set of descriptions in no way precludes certainty about the ranking of the same set of actions under a different set of descriptions. . . . That every action falls under infinite descriptions gives us a fair bit of room to work here" (2009, 23). The fact that one can make this sort of comparison through analogical reasoning is an important feature of the background ranking procedure.

One might reasonably wonder where "fifty" (in 50W) came from. We have admittedly imputed this belief to Tegan in an ad hoc manner. However, as we noted, given that the source of uncertainty for Tegan is regarding the decision to program the car to first crash into the roadster or not, we think it is indeed plausible that Tegan may have a good sense of how the other analogies we have introduced fare against each other.

These three steps capture the essence of the background ranking procedure. Now we are in a position to perform an expected moral value calculation:

(1) Tegan's credence in the proposition, "I should program the autonomous vehicle to crash into the roadster" [*insert value*] × the difference in the magnitude of the moral value between crashing into the roadster and not crashing into the roadster, on the condition that the charity analogy holds [*insert value*]
(2) Tegan's credence in the proposition, "I should not program the autonomous vehicle to crash into the roadster" [*insert value*] × the difference in the magnitude of the moral value between crashing into the roadster and not crashing into the roadster, on the condition that the organ extraction analogy holds [*insert value*]

which is

(1) $(0.9)(W) = 0.9W$
(2) $(0.1)(50W) = 5W$

Finally, to determine what Tegan should do, we simply take the difference between the expected value of programming the vehicle to crash into the roadster (.9W) versus not programming the vehicle to do so (5W). When the value is positive, she should program the vehicle to crash into the roadster, and when it

is negative, she should not. (If the difference is zero, she could justifiably choose either.) Given that .9W−5W is a negative value (−4.1W), from Tegan's perspective, the appropriate choice is "I should not program the autonomous vehicle to crash into the roadster."

We recognize that the proposal we have just articulated has problems (some readers might think serious problems). For instance, some readers may find it objectionable that the procedure does not always prevent one from achieving the intuitively wrong outcomes. This unseemly feature is inherent to any rational decision procedure that incorporates one's subjective inputs. However, it is important to note that we do not claim that the expected moral value approach guarantees the moral truth from the view of objective reason. We aim only to show that the expected value approach offers rational decision-making guidance when the decision maker must make a decision in her current doxastic state.

Perhaps most worrisome is that the procedure on offer might strike some as question-begging. That is, some readers might think that it presupposes the truth of consequentialism. But this is not quite right for two reasons. First, an assumption underlying this worry is that a view that tells in favor of "numbers mattering," as it were, must be a consequentialist view. This assumption is untenable. In the trolley problem, for instance, a Kantian can coherently choose to save more lives because she believes doing so is the best she can do with respect to her deontological obligations (Hsieh, Strudler, and Wasserman 2006). Likewise, the fact that the procedure we defend employs an expected value approach need not mean it is consequentialist. It can be defended on deontological grounds as well.

More fundamentally, the procedure we defend is a meta-normative account that is indifferent to the truth-value of any particular first-order moral theory. This brings us to the second reason the question-begging worry is not problematic.

The approach we defend concerns what the programmer *all things considered* ought to do—what the programmer has strongest reason to do. This sort of all-things-considered judgment of practical rationality incorporates various types of reasons, not exclusively moral ones.

A range of reasons impact what one has strongest reason to do: self-interested, agent-relative, impartial moral, hedonistic, and so on. The fact that moral reasons favor Φ-ing does not settle the question of whether Φ-ing is what one ought to do all things considered. As Derek Parfit notes:

> When we ask whether certain facts give us morally decisive reasons not to act in some way, we are asking whether these facts are enough to make this act wrong. We can then ask the further question whether these morally decisive reasons are decisive all things considered, by outweighing any conflicting non-moral reasons. (Forthcoming)

What practical rationality requires one to do will turn on a range of normative reasons including moral ones, but not only moral ones. It would indeed be worrisome, and potentially question-begging, if we were claiming that the procedure provides the programmer with morally decisive guidance. But this is not what we are doing. We are concerned with helping the programmer determine what the balance of reasons favors doing—what the programmer has strongest reason to do—and this kind of judgment of practical rationality turns on a range of different kinds of reasons (Sepielli 2009).

In sum, what practical rationality requires of the programmer is distinct from what the programmer ought to believe is wrong, and even what the programmer has most moral reason to do. The procedure on offer helps the programmer reason *about* her moral beliefs, and this in itself is not a moral judgment (though what the programmer ought to do all things considered could, of course, coincide with what she has most moral reason to do). The procedure we are defending can help the programmer figure out what the balance of reasons favors doing, what she has *strongest* reason to do all things considered.[11]

1.4 A Brief Moral Defense and Remaining Moral Objections

While the procedure itself concerns practical rationality, we nevertheless think that there may be good *independent* moral reasons that favor the programmer's using such a procedure in the face of moral uncertainty. First, an expected value approach can help a programmer to avoid acting in a morally callous or indifferent manner (or at least to minimize the impact of moral indifference). If the programmer uses the expected value procedure but arrives at the immoral choice, she is less blameworthy than had she arrived at the immoral choice without using the procedure. This is because using the procedure displays a concern for the importance of morality.

If Tegan, in the grip of moral uncertainty, were to fail to use the expected value procedure, she would display a callous disregard for increasing the risk of wronging other human beings.[12] As David Enoch says, "[I]f I have a way to minimize the risk of my wronging people, and if there are no other relevant costs . . . why on earth wouldn't I minimize this risk?" (2014, 241). A programmer who in the face of moral uncertainty chooses to make a decision on a whim acts recklessly. Using the expected moral value approach lowers the risk of wronging others.

A second moral reason for using an expected value approach to deciding under moral uncertainty is that it can help the programmer embody the virtue of humility (e.g., Snow 1995). Indeed, as most professional ethicists admit, moral matters are deeply complex. A programmer who in the grip of moral uncertainty insists on using the "My Favorite Theory" approach fails to respect the difficulty

of moral decision making and thereby exhibits a sort of intellectual chauvinism. Karen Jones's remark seems apt: "If it is so very bad to make a moral mistake, then it would take astonishing arrogance to suppose that this supports a do-it-yourself approach" (1999, 66–67). The fact that a programmer must take into account the possibility that she is mistaken about her moral beliefs builds a kind of humility into the very decision procedure.

Some may worry that a programmer who uses the expected moral value approach is compromising her integrity (Sepielli, forthcoming). The thought is that the programmer who follows the approach will often have to act in accordance with a normative view that she thinks is less likely correct. And acting in accordance with a theory one thinks less likely to be correct compromises one's integrity. Though this line of thought is surely tempting, it misses its mark. The value of integrity is something that must be considered along with other moral values. And how to handle the importance of the value of integrity is again a question that may fall victim to moral uncertainty. So the issue of integrity is not so much an objection as it is another consideration that must be included in the set of issues a programmer is morally uncertain about.

Another objection to a programmer using an expected value procedure holds that the programmer would forgo something of great moral importance—that is, moral understanding. For instance, Alison Hills (2009) claims that it is not enough to merely make the right moral judgments; one must secure moral understanding.[13] She argues that even reciting memorized reasons for the right actions will not suffice. Instead, she claims, one must develop understanding—that is, roughly, the ability to synthesize the moral concepts and apply the concepts in other similar contexts. And clearly, a programmer who is inputting her credences and related normative beliefs into an expected moral value procedure lacks the sort of moral understanding that Hills requires.

But this sort of objection misses the point. First, it is important to keep in mind that while the outcome of the procedure might not have been what a programmer originally intended, it is the programmer herself who is deciding to use the procedure that forces her to consider the moral implications of the possibility of deciding incorrectly. Second, it would indeed be the ideal situation to develop moral understanding, fully exercise one's autonomy, and perform the action that the true moral view requires. However, we agree with Enoch, who aptly notes:

> Tolerating a greater risk of wronging others merely for the value of moral autonomy and understanding is thus self-defeating, indeed perhaps even practically inconsistent. Someone willing to tolerate a greater risk of acting impermissibly merely in order to work on her (or anyone else's) moral understanding, that is, independently of the supposed instrumental payoffs of having more morally understanding people around, is acting

> wrongly, and indeed exhibits severe shortage in moral understanding (of the value of moral understanding, among other things). (2014, 249)

One can imagine how absurd an explanation from a programmer who decided on her own and acted in a way that wronged someone would sound if she were asked by the person who was wronged why she did not attempt to lower the risk of wronging another and she responded, "Well, I wanted to exercise my personal autonomy and work on my moral understanding."[14] Such a response would be patently offensive to the person who was wronged given that the programmer did have a way to lower her risk of wronging another.

1.5 Conclusion

In this chapter we aimed to show that programmers (or whoever will ultimately be choosing the ethics of autonomous vehicles) are likely to face instances where they are in the grip of moral uncertainty and require a method to help them decide how to appropriately act. We discussed three proposals for coping with this uncertainty: Continue Deliberating, My Favorite Theory, and a particular expected moral value approach. We offered some moral reasons for why the programmer has reasons to employ the third procedure in situations with moral uncertainty.

While there are surely many remaining issues to be discussed with respect to the question of how to deal with moral uncertainty in programming contexts, this chapter aims to provide a first step to offering programmers direction on how to appropriately handle decision-making under moral uncertainty. We hope it encourages robot ethics scholars to pay more attention to guiding programmers who are under moral uncertainty.

Notes

1. Patrick Lin (2014, 2015) is one of the first scholars to explore the relevance of the trolley problem in the context of autonomous vehicles.
2. There are important moral questions that we do not consider in this chapter. For instance, should the age of passengers in a vehicle be taken into account in deciding how the vehicle should be programmed to crash? Who should be responsible for an accident caused by autonomous vehicles? Is it possible to confer legal personhood on the autonomous vehicle? What liability rules should we as society adopt to regulate autonomous vehicles? (Douma and Palodichuk 2012; Gurney 2013). For ethical issues involving robots in general, see Nourbakhsh (2013), Lin, Abney, and Bekey (2012), and Wallach and Allen (2009).
3. Robotics Institute of Carnegie Mellon University, for instance, offers a course named "Ethics and Robotics."

4. Here we mirror the formulation of an example in MacAskill (forthcoming).
5. As will become clear, we owe a significant intellectual debt to Andrew Sepielli for his work on the topic of moral uncertainty.
6. One might think that technological advances can soon minimize empirical uncertainties. But this is a naive assumption. Existing robots are far from being able to fully eliminate or account for possible empirical uncertainties. We are grateful to Illah Nourbakhsh, professor of robotics at Carnegie Mellon University, for bringing this point to our attention.
7. The solution we support resembles the celebrated *Pascal's Wager* argument that Blaise Pascal offers for why one should believe in God. Pascal states, "Either God is or he is not. But to which view shall we be inclined? . . . Let us weigh up the gain and the loss involved in calling heads that God exists. Let us assess two cases: if you win you win everything, if you lose you lose nothing. Do not hesitate then; wager that he does exist" (1670, § 233). We recognize that many find Pascal's wager problematic (Duff 1986), although there are writers who find it logically valid (Hájek 2012; Mackie 1982; Rescher 1985). At any rate, we are not concerned here to intervene in a debate about belief in divine existence. Nevertheless, we do think insights from Pascal's Wager can usefully be deployed, with relevant modifications, for the problem of programming under moral uncertainty.
8. Sepielli considers a scenario much like the one Tegan is in. He notes, "Some consequentialist theory may say that it's better to kill 1 person to save 5 people than it is to spare that person and allow the 5 people to die. A deontological theory may say the opposite. But it is not as though the consequentialist theory has, somehow encoded within it, information about how its own difference in value between these two actions compares to the difference in value between them according to deontology" (2009, 12).
9. Philosopher Ted Lockhart offers a proposal that aims to hedge and also claims to avoid the PIVC. Lockhart's view requires one to maximize "expected moral rightness" (Lockhart 2000, 27; Sepielli 2006) and thus does indeed account not only for the probability that a particular moral theory is right, but also for the moral weight (value, or degree) of the theory. One important problem with Lockhart's view is that it regards moral theories as having equal rightness in every case (Sepielli 2006, 602). For a more detailed criticism of Lockhart's position, see Sepielli (2006, 2013).
10. The example we use to explain the three steps also closely models an example from Sepielli (2009). It is worth noting that Sepielli does not break down his analysis into steps as we have. We have offered these steps with the hope that they accurately capture Sepielli's important insights while also allowing for practical application.
11. We are grateful to Suneal Bedi for helpful discussion regarding the issues in this section.
12. David Enoch (2014) offers this reason for why one ought to defer to a moral expert with regard to moral decisions.

13. Hills states, "Moral understanding is important not just because it is a means to acting rightly or reliably, though it is. Nor is it important only because it is relevant to the evaluations of an agent's character. It is essential to acting well" (2009, 119).
14. This sort of objection to Hills (2009) is due to Enoch (2014). Enoch objects in the context of a person failing to defer to a moral expert for moral guidance.

Works Cited

De Groot, Morris H. 2004. *Optimal Statistical Decisions*. New York: Wiley-Interscience.

Douma, Frank and Sarah A. Palodichuk. 2012. "Criminal Liability Issues Created by Autonomous Vehicles." *Santa Clara Law Review* 52: 1157–69.

Duff, Anthony. 1986. "Pascal's Wager and Infinite Utilities." *Analysis* 46: 107–9.

Enoch, David. 2014. "A Defense of Moral Deference." *Journal of Philosophy* 111: 229–58.

Fagnant, Daniel and Kara M. Kockelman. 2013. *Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations*. Washington, DC: Eno Center for Transportation.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5: 5–15.

Goodall, Noah J. 2014a. "Ethical Decision Making During Automated Vehicle Crashes." *Transportation Research Record: Journal of the Transportation Research Board*, 58–65.

Goodall, Noah J. 2014b. "Vehicle Automation and the Duty to Act." *Proceedings of the 21st World Congress on Intelligent Transport Systems*. Detroit.

Gurney, Jeffrey K. 2013. "Sue My Car Not Me: Products Liability and Accidents Involving Autonomous Vehicles." *Journal of Law, Technology & Policy* 2: 247–77.

Gustafsson, John. E. and Olle Torpman. 2014. "In Defence of My Favourite Theory." *Pacific Philosophical Quarterly* 95: 159–74.

Hájek, Alan. 2012. "Pascal's Wager." *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/pascal-wager/.

Hare, Caspar. 2012. "Obligations to Merely Statistical People." *Journal of Philosophy* 109: 378–90.

Harman, Elizabeth. 2015. "The Irrelevance of Moral Uncertainty." In *Oxford Studies in Metaethics*, vol. 10, edited by Luss-Shafer Laundau, 53–79. New York: Oxford University Press.

Hills, Alison. 2009. "Moral Testimony and Moral Epistemology." *Ethics* 120: 94–127.

Hsieh, Nien-ĥe, Alan Strudler, and David Wasserman. 2006. "The Numbers Problem." *Philosophy & Public Affairs* 34: 352–72.

Jones, Karen. 1999. "Second-Hand Moral Knowledge." *Journal of Philosophy* 96: 55–78.

Lin, Patrick. 2014. "The Robot Car of Tomorrow May Just Be Programmed to Hit You." *Wired*, May 6. http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/

Lin, Patrick. 2015. "Why Ethics Matters for Autonomous Cars." In *Autonomes Fahren*, edited by M. Maurer, C. Gerdes, B. Lenz, and H. Winner, 70–85. Berlin: Springer.

Lin, Patrick, Keith Abney, and George A. Bekey, eds. 2012. *Robot Ethics: The Ethical and Social Implications of Robotics*. Cambridge, MA: MIT Press.

Lockhart, Ted. 2000. *Moral Uncertainty and Its Consequences*. New York: Oxford University Press.

MacAskill, William. Forthcoming. "Normative Uncertainty as a Voting Problem." *Mind*.

Mackie, J. L. 1982. *The Miracle of Theism*. New York: Oxford University Press.

Millar, Jason. 2014. "You Should Have a Say in Your Robot Car's Code of Ethics." *Wired*, September 2. http://www.wired.com/2014/09/set-the-ethics-robot-car/.

Nourbakhsh, Illah R. 2013. *Robot Futures*. Cambridge, MA: MIT Press.

Parfit, Derek. Forthcoming. *On What Matters*, part 3. Oxford University Press.

Pascal, Blaise. (1670) 1966. *Pensées*. Translated by A. K. Krailsheimer. Reprint, Baltimore: Penguin Books.

Raiffa, Howard. 1997. *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. New York: McGraw-Hill College.

Rescher, Nicholas. 1985. *Pascal's Wager*. South Bend, IN: Notre Dame University.

Sepielli, Andrew. 2006. "Review of Ted Lockhart's *Moral Uncertainty and Its Consequences*." *Ethics* 116: 601–4.

Sepielli, Andrew. 2009. "What to Do When You Don't Know What to Do." In *Oxford Studies in Metaethics*, vol. 4, edited by Russ-Shafer Laundau, 5–28. New York: Oxford University Press.

Sepielli, Andrew. 2010. "Along an Imperfectly-Lighted Path: Practical Rationality and Normative Uncertainty." PhD dissertation, Rutgers University.

Sepielli, Andrew. 2012. "Normative Uncertainty for Non-Cognitivists." *Philosophical Studies* 160: 191–207.

Sepielli, Andrew. 2013. "What to Do When You Don't Know What to Do When You Don't Know What to Do . . . ," *Nous* 47: 521–44.

Sepielli, Andrew. Forthcoming. "Moral Uncertainty." In *Routledge Handbook of Moral Epistemology*, edited by Karen Jones. Abingdon: Routledge.

Snow, Nancy E. 1995. "Humility." *Journal of Value Inquiry* 29: 203–16.

Thomson, Judith Jarvis. 1976. "Killing, Letting Die, and the Trolley Problem." *Monist* 59: 204–17.

Urmson, Chris. 2014. "Just Press Go: Designing a Self-driving Vehicle." Google Official Blog, May 27. http://googleblog.blogspot.com/2014/05/just-press-go-designing-self-driving.html.

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

Werhane, Patricia. 1999. *Moral Imagination and Management Decision-Making*. New York: Oxford University Press.