#  11

# Autonomous Vehicles and Ethical Settings

## Who Should Decide?

*Paul Formosa*

## Introduction

Autonomous vehicles (AVs) will be placed in difficult moral scenarios where they will be forced to choose between two bad outcomes, such as the death of a pedestrian or a passenger. While, unlike autonomous military weapons systems or "killer robots" (Smith 2019), AVs are not *designed* to harm people, harming people is an inevitable *by-product* of their operation. How are AVs to deal ethically with situations where harming people is inevitable? Rather than focus on the much-discussed question of *what* choices AVs should make in such cases (Gerdes and Thornton 2016; Lin 2016; Nyholm 2018; Scheutz 2016), we can also ask the much less discussed question of *who* gets to decide what AVs should do in such cases (Millar 2017). Here there are two key options (Gogoll and Müller 2017): AVs with a personal ethics setting (PES) also known as an "ethical knob" (Contissa, Lagioia, and Sartor 2017) that end users can control, or AVs with a mandatory ethics setting (MES) that end users cannot control. Which option, a PES or an MES, is best and why? In this chapter I argue, drawing on the choice architecture literature (Thaler, Sunstein, and Balz 2012), in favor of a hybrid view that requires mandated default choice settings while allowing for limited end user control.

## The Moral Argument for AVs

The moral argument for AVs is straightforward: they could potentially save a lot of lives.[1] If we have a new piece of technology that could save a lot of lives, then this gives us prima facie reasons to introduce it as soon as we safely can.[2] But how many lives might AVs save? The World Health Organization's (2018) most recent

Paul Formosa, *Autonomous Vehicles and Ethical Settings* In: *Autonomous Vehicle Ethics*. Edited by: Ryan Jenkins, David Černý, and Tomáš Hříbek, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197639191.003.0011

report puts the number of deaths from road traffic crashes at 1.35 million a year in 2016, with additional tens of millions of people injured or disabled every year from traffic accidents. Globally, road traffic injury is the leading cause of death for people aged between five and twenty-nine years and is also the eighth most common cause of all deaths, killing more people per year than HIV/AIDS, tuberculosis, or diarrheal diseases. However, many of those global deaths occur in low-income countries where the death rates from traffic injuries are three times higher than in high-income countries. Given the high initial costs of AV technology, it may be some time before AVs can do much to help the global poor. But even if we limit our initial focus to high-income countries, the numbers of lives that could be saved by AVs are still very large. For example, in 2013 in the United States about 32,000 people died due to traffic accidents and roughly 93 percent of the 5.5 million traffic crashes that occurred that year have been attributed to human error, such as drunk driving, speeding, distraction, fatigue, and poor driving skills (Gogoll and Müller 2017). But AVs won't get drunk, distracted, or fatigued, and this has led many to believe that AVs will *significantly* reduce the overall number of traffic injuries (Gogoll and Müller 2017). Many tens of thousands of lives could be saved, and many millions of injuries avoided, *every single year* through the introduction of AVs in high-income countries alone. Those numbers promise to be many times greater as AVs become widespread in lower-income countries. AVs also promise to bring significant environmental and economic benefits, even though they raise privacy concerns (Wolkenstein 2018). Overall, the moral case for AVs is clearly very strong, assuming they can live up to their promise.[3]

Nonetheless, AVs will not eliminate all traffic injuries and deaths. While some of these injuries might be caused by software or hardware failures, the cases we shall focus on here are due to the presence of *tragic scenarios* where someone will inevitably be harmed by an AV. One such example is the *tunnel case* (Gogoll and Müller 2017; Millar 2015, 2017). In this case a school bus in front of you breaks suddenly. If your AV applies the brakes and doesn't swerve, your AV will kill three children on the bus, but you will survive. If your AV swerves left, you push the car next to you into the tunnel wall, killing the two people inside, but you will survive. If your AV swerves right, you run into the tunnel wall and die, but no one else is injured. With many AVs traveling many kilometers, variations of such tragic scenarios will arise repeatedly (Jenkins 2016; Wolkenstein 2018). Even the safest AVs will still kill and injure some people, even if they save many lives compared to an otherwise equivalent world with human drivers and no AVs. Such cases raise important legal issues around liability for the harms caused by AVs (Hevelke and Nida-Rümelin 2015; Santoni de Sio 2017), but rather than look at those issues, we shall instead focus on the *ethical settings* that AVs need to deal with such cases.

## AVs Need Ethical Settings

Why do AVs need ethical settings? Human drivers also face tragic scenarios, but we don't worry about our own ethical settings in such cases. Perhaps we should. In any case, there are several key differences between AVs and humans which explain why AVs need ethical settings (Faulhaber et al. 2019; Gogoll and Müller 2017). Imagine a variation of the *tunnel case* where a human is driving the car in question. What difference does this make? In both cases, there is only a split second to decide what to do. There is no time for a human to consciously deliberate about what they should do while taking relevant moral principles into account. Instead a human must make a sudden instinctual choice, with limited information, while pumped full of adrenalin. In contrast, AVs don't have adrenalin, have much more information at their disposal, and enough time to make a calculated decision about what they should do. When humans kill someone as the result of a tragic split-second choice, such as in the *tunnel case*, it is an *accident*. When an AV does the same thing, the outcome "cannot be considered accidental" (Gogoll and Müller 2017, 686) because it is the *calculated* result of an algorithm. Thus, AVs need an ethical setting to determine what they should do in such cases (Lin 2016; Millar 2017). Simply, as some suggest (van Wynsberghe and Robbins 2019), designing AVs to be as "safe" as possible and leaving the ethics to humans won't work, since in tragic scenarios such as the *tunnel case*, it ceases to be a matter of safety alone because the AV must decide *whose* safety to *prioritize*. And that is a moral question, not a safety question, and one that cannot be offloaded to a human since there is not enough time (Formosa and Ryan 2021). In tragic scenarios, circumstances dictate that AVs *must* make calculated moral choices, and to do that they need ethical settings.

Much of the discussion of *what* such ethical settings should look like has been framed in terms of trolley problems. Putting aside the issues with extrapolating from trolley problems to AVs,[4] a further problem is that framing the discussion in this way tends to oversimplify the options as either a choice between utilitarian or deontological (or Kantian) AVs (Faulhaber et al. 2019). But this focus obfuscates several significant and complex ethical issues that the ethical settings on AVs will need to take a stand on. Since it is important for arguments developed herein, it is necessary that we get a brief taste of that complexity here.

The first of these issues is a setting about the priority of those inside AVs versus those outside AVs. Here there are at least three broad ethical settings (Gogoll and Müller 2017).[5] *Selfish*: weigh the lives of those in the AV above other lives. *Equality*: weigh the lives of those in the AV as equivalent to other lives. *Altruism*: weigh the lives of those in the AV below other lives. This setting has clear real-world implications. In the *tunnel* case, the *selfish* setting would run you into the bus (or the other car) even though this kills three children (or two others), whereas both the *equality* and *altruism* setting would run you into the wall. However, if your AV had three passengers in it, *altruism* would run your AV into the wall, whereas *equality* would swerve left and kill the two people in the car next to you. Further, each of these three broad options allows for many variations. For example, one person's *selfish* setting could weigh their own life as worth twice as much as others, whereas another person's *selfish* setting could set that weight to ten, a hundred, or infinite times more valuable than the lives of others.[6]

The second and related issue is a setting about how to weigh lives in general. This is an issue even if your car has an *equality* setting, since there is more than one way to treat everyone as equal. In the *tunnel case* we are weighing up options with different *person-numbers* (the number of people who will live or die). Now consider another case. In the *oil slick* case, an AV has hit an oil slick and is about to crash into one of two pedestrians, either Amy or Belinda. Who should it hit? What if Amy is a ninety-year-old widow with advanced cancer and Belinda is a healthy three-year-old girl?[7] Or what if Amy is the world's worst criminal and Belinda is the world's greatest philanthropist, or if Amy is the sole caregiver to three young children and Belinda has no children, or if Amy is unemployed and Belinda is the greatest living scientist whose work could save the lives of millions? To respond to these questions, we need to know how morally to distribute a scarce good (i.e., not getting killed by an AV).

For utilitarians, standardly, we should decide this on the basis of what maximizes *total* utility, and only indirectly consider *person-numbers* and the *distribution* of utility insofar as it impacts on total utility. Unfortunately, in the discussion of AVs utilitarianism is often simply associated with minimizing the "number of deaths" (Contissa et al. 2017, 370). But utilitarianism cares *directly* about maximizing *total utility* and that doesn't always equate with maximizing *person-numbers*. For example, if an AV has to choose between killing one person or ten persons, and the one person is a great benefactor of humankind who helps millions and the ten are mean misers who help no one, then the greater total utility might be gained not by minimizing the number of deaths but by killing the ten misers to save the one benefactor.

For Kantians, however, the focus on total utility is a morally inappropriate basis on which to make such a decision, since persons are not mere utility containers but moral agents who individually deserve respect and have dignity. How can Kantians deal with such cases? Kerstein (2013) argues that there are at least three Kantian options here. First, give each person an equal chance of being saved. Second, give each person an equal weighted chance of being saved. Third, maximize person-numbers and "person-years" (the number of years of agency left). To simplify matters, in *modified-tunnel*, the passenger of the AV is always safe, and the choice is between running into the bus and killing two children or running into the other car and killing one child. In *modified-tunnel* that would mean: for the first view, having a ½ chance the AV will run into the bus and a ½ chance it will run into the other car; for the second view, having a 2/3 chance the AV will run into the car and a 1/3 chance that it will run into the bus;[8] and on the third view, crashing into the other car as this maximizes person-numbers. In *oil slick* it would mean: on the first two views, giving a ½ chance of survival to both Amy and Betty; and, on the third option, choosing to kill whoever has less person-years left (i.e., the older and/or sicker of the two).[9] However, we could also imagine many other possible ethical views that take other features into account, such as persons' relative economic productivity, intelligence, achievements, whether they have dependent family members, whether they are your friends or family, and so on. Different ethical settings could thus weigh different factors, and weigh those factors differently, in determining what to do in cases such as *tunnel* or *oil slick*.

More briefly, a third issue is a setting about who counts as a person when we are counting person-numbers for the purposes of moral calculations by AVs (Formosa 2017). For example, does a heavily pregnant woman count as one or two persons? And how far along must the pregnancy be? Does an ape who knows sign language or a human in a permanent vegetative state count as a person? A fourth issue is a setting about whether we should make a moral distinction between those "involved" who can be part of the AV's moral calculations, such as other road users, and those who are "uninvolved" and cannot be used in this way, such as people enjoying coffee on a café pavement (Hübner and White 2018). A fifth issue is a setting about how to weigh up various nonfatal harms to different persons. For example, how should an AV weigh up crushing one person's foot with crushing another person's hand? Or weigh up breaking the legs of ten people or making another person blind? A sixth issue is a setting about how to weigh harms to persons and harm/damage to nonpersons. For example, how should an AV weigh up slightly injuring one person or killing ten healthy dogs (assuming dogs are not persons)? A seventh issue is a setting about how to weigh up harm/damage to different nonpersons. For example, how should an AV weigh up running into a fence or knocking over a very old tree? An eighth issue is a setting about how to deal with risk, given that AVs will be dealing with probabilities and not certainties (Contissa et al. 2017; Nyholm and Smids 2016). How should an AV weigh up a 99 percent risk of damage to property with a 0.0001 percent risk of very minor harm to a person's foot? Or a 10 percent risk of death to one person with an 85 percent risk of serious but nonfatal harm to ten people?

Each of these several ethical issues needs some sort of setting, and each clearly has many plausible variants. To be comprehensive and thereby avoid leaving ethically significant issues merely implicit and unexamined, an ethical setting in an AV will need to deal explicitly with all these (and likely many other) moral issues. But an ethical setting that deals with *all* these issues, as well as all the possible combinations of different options, will have to be *very* complex. Rather than worry here about the content of that complex ethical setting, we shall instead consider the question of *who* should decide what that ethical setting should be. Here we have two main options (Gogoll and Müller 2017): a personal ethics setting (PES) where end users have control over ethical settings, and a mandatory ethics setting (MES) where they do not have that control.[10] Which of these is the best option?[11]

## Arguments for and against a PES and an MES

The most important argument in favor of a PES is the *popularity argument*, which says that we should have a PES because people strongly prefer it. People say they are more likely to buy or use an AV that puts their safety above others (a *selfish* setting), but they prefer that others not adopt such a setting (an *equality* setting).[12] If people won't buy or use AVs that they can't set to have a *selfish* setting, then uptake of AVs will be much lower, and the moral benefits of AVs (namely, many lives saved) won't be fully realized. This tells us that what people really want is to freeride by adopting a *selfish* setting while others adopt an *equality* setting. However, it seems unlikely that a PES would deliver that outcome since, as we shall see later, it is rational for everyone to adopt a selfish setting, leaving everyone worse off (Gogoll and Müller 2017). In any case, the force of the *popularity argument* is premised on the claim that, at least initially, there will *only* be a large uptake of AVs if there is a PES. However, once people see the benefits of AVs for themselves, this attitude might change. Further, the strength of this argument might also depend on the model by which AVs become widespread. There are at least two options here: a personal ownership model and a taxi/rideshare model. If the latter model becomes the dominant one, then people might not expect the option of a *selfish* ethical setting since people probably won't expect that level of control in a taxi/rideshare service, as opposed to in AVs they personally own and would therefore expect more control over.[13] If the taxi/rideshare model becomes the dominant one, then the lack of a PES might be less of an issue.

A related argument is the *respect* (or *autonomy*) *argument*. We should respect people's autonomy by allowing them to make their own choices in morally important areas, especially when the moral stakes are high (Millar 2017). Who your AV is programmed to kill in tragic scenarios is a morally important matter, and thus we should let people make their own choices here by giving them a PES (Jenkins 2016). But against this, it is a standard liberal claim, encapsulated in Mill's harm principle (Turner 2014), that we have no right to choose to harm others. Since we are dealing with ethical settings that have to do with harming others, the respect argument does not seem to apply here. Indeed, the fact that we are dealing with calculated harms to others suggests, as the later *justice argument* makes explicit, that the ethical settings in AVs should be a matter of common regulation (i.e., MES) and not individual choice (i.e., PES).

This point feeds into the *disagreement argument*. As we have seen, there are lots of possible ethical settings. Which should we choose? Clearly there will be disagreement about this. If we mandate an MES, we must be able to justify it. But can we do that, in the face of the inevitable disagreement? One way to deal with disagreement is to let people individually choose for themselves, in this case by giving them a PES (Gogoll and Müller 2017). But we don't outsource to individuals the choice about what to do *whenever* there is widespread disagreement. For example, we disagree about how much taxation people should pay, but we don't let everyone decide individually how much tax to pay. Indeed, in almost *any* area that governments make decisions, there is often sizeable disagreement, and yet governments don't abdicate every such decision to individuals. There is no reason why it shouldn't be the same with ethical settings in AVs.

The two most important arguments against a PES are the *bad choices* and the *complexity arguments*. The *bad choices argument* is that people might make horrible choices, such as opting for racist or sexist settings (Gogoll and Müller 2017). For example, if we were to allow *any* setting, one could imagine an AV that sought to maximize a racist's conception of the good by intentionally killing black people, or at least counting their lives as worth less than a white person's life. One response is to limit the options to a predetermined list that rules out any grossly biased options (Gogoll and Müller 2017; Contissa et al. 2017; Millar 2014). But this response leads into the next worry. The *complexity argument* says that the choice of an ethical setting for an AV is too complex for most people to be able to make an informed choice about it (Millar 2017). Consider all the issues we mentioned earlier about the *content* of ethical settings and all the different permutations of the various options discussed. For example, one might prefer an ethical setting that has a selfish preference that weighs your life as worth exactly three times that of others, seeks to maximize agent years and not happiness, takes large risks, values property highly, does not care about animals, holds that apes count as persons, that those on pavements are uninvolved and should not be hit, and that those with a criminal record for crimes with a prison sentence greater than three years can be killed first. Clearly, there are a *lot* of options here, and explaining them all in sufficient detail would be very complex. One response to the complexity problem is the same as to the previous worry, namely to use a predetermined list with a short number of simple options. But it is not clear that this could lead to *informed* choice. Consider the related "transparency paradox" which is discussed in the context of giving consent online to privacy conditions (Nissenbaum 2011, 36). Either what one is consenting to is too simplified and so one cannot give proper informed consent, or it is too complex to understand and so once again one cannot give proper informed consent. A similar concern applies here. Either the ethical settings are too few and too simple, in which case much is being left out that is ethically important, and so one cannot make a fully informed choice; or the settings are too many, too complex, and too detailed, in which case no one (or almost no one) can really understand them, and so again one cannot make a fully informed choice.[14]

The main argument in favor of an MES is the *justice argument*, which says that serious calculated harms to others are collective political or justice issues requiring mandated solutions, not personal ethical ones to be left up to each individual to decide. Since the ethical settings on AVs are matters of nonaccidental serious harms to other persons, this is a domain of justice or politics (i.e., an MES), and not personal ethics (i.e., a PES) (Himmelreich 2018). Further, a just and fair MES should be determined, not by ad hoc industry standardization or market forces, but through a collective process of fair, open, and democratic decision-making that no one could reasonably reject (Hübner and White 2018; cf. Millar 2017). This process will help to ensure that whatever MES is adopted is widely seen as justified. An MES will also help to address complexity problems, since complex decisions, such as the ethical settings in AVs, are precisely the sort that lend themselves to formal public deliberation where experts can collectively help to deal with the complexity that prevents most individuals from being able to make informed choices.

The main arguments against an MES are the flipsides of the arguments for a PES already discussed earlier. The most important of these is the flipside of the *popularity argument*, which is that an MES will significantly stifle the uptake of AVs as it will be unpopular. A further argument against an MES is that it is too limiting of personal choice. In particular, while we might not want to mandate altruism, it should be *permissible* for people to opt for an altruistic AV. But an MES would seem to remove any such option. One response to this worry is that an MES could include an "altruistic add-on" (Gogoll and Müller 2017, 698). But then AVs would need a limited PES to turn on the optional altruism setting. This suggests that perhaps a hybrid view, which mixes elements of a PES and an MES, might be a better overall option. We explore such an option in the next section.

## Choice Architecture and AVs

As we have seen, the most significant worry with an MES is that it will stifle the uptake of AVs as people won't buy or use them, and one of the most significant worries with a PES is that it will lead to worse overall outcomes as everyone will end up choosing selfish AVs. Is there a way to get the best of both worlds while avoiding the worst of each? I shall argue, by drawing on the choice architecture literature, that there is. To see why this is, we first need to look at Gogoll and Müller's (2017) argument against a PES in more detail. They argue that a PES will lead to a prisoner's dilemma outcome where everyone acting in their own self-interest results in everybody being worse off than if they cooperated. While the technical details are not important here, the main thrust of the argument is straightforward. In a world of AVs with a PES, the rational thing to do is to adopt a *selfish* setting. But since everyone reasons in the same way, and no one wants to be the only person with an *equality* or *altruism* setting, everyone will rationally adopt a *selfish* setting. But a world where everyone adopts a *selfish* setting is worse than a world in which everyone cooperates by adopting an *equality* setting since the latter leads to less overall harm than the former. Is this a good argument against a PES?

In response, it is unclear whether *everyone* will be worse off under a universally adopted selfish PES, since even if this results in more deaths *overall*, the *distribution* of harm matters, and *some* people might be better off under this distribution. In any case, this is only a problem if in the real world everyone acts as a perfectly rational agent when it comes to their PES. But we know that people are not perfect rational maximizers. Indeed, this point is central to the entire choice architecture literature. Choice architecture is about "organizing the context in which people make decisions" in order to "indirectly influence [or "nudge"] the choices" people make (Thaler et al. 2012, 428). One of the key principles of choice architecture theory is that people choose the path of least resistance, and if there is a default choice that requires them to do nothing, then most people will end up with that option "whether or not it is good [i.e., rational] for them" (Thaler et al. 2012, 430). This is contrasted with a required choice architecture where there is no default option. There are at least two conditions where we should prefer a default choice over a required choice architecture. First, when a "choice is complicated and difficult, people might greatly appreciate a sensible default." Second, when "choices are highly complex, required choosing may not be a good idea; it might not even be feasible" (Thaler et al. 2012, 431). Both these conditions are met here. Requiring people to make a choice about ethical settings before they can use an AV is requiring them to make a very complicated, difficult, and complex decision about matters which they are not likely to be properly informed about. In such cases, a default choice architecture is most appropriate.

Beyond the appropriateness of a default choice, the choice architecture literature provides other relevant guidance (Thaler et al. 2012, 433–35). First, "Give feedback." After an ethical setting has been utilized, the AV could give feedback on who was saved and why. Second, provide "mappings" from "choice to welfare." There should be clear explanations about the real-world impacts different settings will have. For example, the choice context could make clear that if you choose a *selfish* setting and your AV must choose between harming you and killing ten children, it will choose to kill the ten children. It could then ask: Are you sure you want it to do that? What about the lives of the children? This could nudge people toward an *equality* setting. Third, “Structure complex choices.” When dealing with large and complicated choice sets, people tend not to carefully weigh up the trade-offs between the alternatives and instead use simplifying algorithms. Default settings are important aides in this context.

If we have a PES with *required choice* (i.e., people *must* select an option from a list before the AV works), then many people might pick the *selfish* setting, which would lead to the suboptimal outcome that Gogoll and Müller use to argue against a PES. But if we instead have a PES with a *mandated default choice* (i.e., no choice is required and instead a mandated default is automatically selected), and use other choice architecture design features to nudge people toward whatever setting is judged to be best (such as an *equality* setting), then many people will keep that default setting even if it is not rational for them to do so. This means that we can nudge most people toward, for example, an *equality* setting, even though they *could* still select a *selfish* setting. This means that we can avoid the bad outcomes associated with a pure PES (i.e., everyone with a *selfish* setting) without having to give up on a PES altogether, since most people will not adopt a *selfish* setting (even if it is rational for them to do so) where a default *equality* setting (or whatever setting is judged to be best) is in place.

This hybrid approach of a mandated default choice architecture promises to give us the best of both PES and MES worlds. Since, under this approach, people *can* change their AV’s ethical settings, including by opting for a *selfish* mode, it is likely to lead to the strong uptake and acceptance of AVs associated with a pure PES. But we also get the advantages of an MES since most people will keep the mandated default choice settings that we collectively judge to be best, thereby avoiding a race to the selfish bottom that might follow from a PES with a required choice architecture. While this hybrid approach blurs the lines between an MES and a PES, those lines were already blurry. For example, any workable PES could only ever offer an unrealistically simplistic choice set in two senses. First, for practical reasons there could only ever be a few options offered, and certain options, such as racist settings, might be completely forbidden. This is equivalent to mandating that only certain options are allowed. Second, in terms of the transparency paradox, either the setting and its description will be too simplistic to cover the relevant detail, or it won’t be, in which case it will be too complex for anyone to properly understand (except, perhaps, for a few experts). In both senses, informed personal choice is already restricted to a degree even under a pure PES. Requiring a mandated default choice is merely more of the same, rather than a radical departure.

There are at least two important worries with this hybrid approach. The first is that, since there is a mandatory default setting, it must be justified. But we will have disagreement about what the setting should be. How can we solve this? As noted earlier, disagreement alone doesn't mean that we should have no regulation. Even so, this hybrid approach is in a better position than a pure MES when it comes to meeting this challenge since, while a mandatory default setting requires some justification, it bears a considerably lower justificatory burden than a pure MES. This is because a mandatory default setting merely *nudges* people in a certain direction, it doesn't (unlike an MES) *lock* people into a setting that they can't override, and so it bears a lesser justificatory burden. We should look to meet this less onerous justificatory burden through a fair and open process, informed by expert opinion, that results in democratic choice that we can all see as reasonable, even if we personally disagree. This collective decision-making machinery could also help to alleviate the complexity concerns that plague individuals trying to grapple with such details. Further, a PES cannot completely avoid the justificatory burden problem either, since it will inevitably have to limit the choice set offered to individuals in terms of both the number of options given and the simplistic description it gives of those options, and both these restrictions require justification. The second worry is that, as the *justice argument* points out, nonaccidental serious harm to others is normally a matter of collective justice and not personal ethics. Why should we treat the ethical settings on AVs differently? Recall that the moral argument for AVs is that they will save a lot of lives and prevent a lot of injuries. This can only happen to its full extent if there is widespread uptake and acceptance of AVs. AVs with a PES, even one with a default choice, should help to ensure that good outcome. But on the hybrid approach advocated here we are not leaving matters *completely* up to individual choice, because we have collectively chosen a mandated default setting which we know most people will keep. Further, we can concede that, if in the future AVs gain very high levels of public trust and consequently the widespread acceptance of AVs ceases to require a PES of any sorts, then there may no longer be a good reason to retain a PES (even one with a default choice), given the strength of the *justice argument*.

This last point suggests the possibility of a two-stage process. The first stage is the hybrid approach advocated here of a PES with a mandated default choice to drive the initial acceptance and uptake of AVs, without most of the negative consequences of a PES with required choice. The second stage, if the acceptance and uptake of AVs becomes firmly established even without a PES, is a pure MES (or, perhaps, a much more restricted PES) that does not allow individuals to opt for their own ethical settings, such as a *selfish* setting. There are at least two reasons to prefer such a two-staged process over a move straight to an MES. The first is that a PES with a default setting is much more likely than an MES to promote strong *initial* support and uptake of AVs and, since this is the moral point of

AVs, this gives us reason to prefer this option *initially*. The second is that, as noted earlier, the justificatory burden that a PES with a default choice needs to meet is far less considerable than those an MES must meet. Given that the ethics of AVs is a relatively new phenomena, the complexity involved is immense, and the real-world practical implications of different options are unclear, it might be hard for any MES to meet those justificatory burdens, at least initially. By starting with an MES we might fail to see alternatives that could have been better, whereas a PES with a default choice allows more room for different options to emerge.[15] Thus, we can accept that a PES with a default choice is the best initial option, while also accepting that, in the longer term, a move to an MES might be required by the *justice argument* once the widespread acceptance of AVs no longer depends on the presence of a PES.

## Conclusion

The debate about who gets to determine the ethical settings in AVs has been cast as one between a PES or an MES. But the line between these two options is blurrier than it seems at first sight. Importantly, there is also a third option, namely a hybrid approach that involves a PES with a mandated default choice. Since this approach allows individual choice, it should encourage the uptake and acceptance of AVs. But since most users won't exercise that choice if we adopt good choice architecture design and mandate a default setting, this approach will nudge most people into accepting a collectively mandated ethical setting. This allows us to get the moral benefits of AVs through their wide acceptance, while avoiding the moral costs of most people opting for selfish settings that make everyone comparatively worse off. However, given that we don't normally leave the regulation of nonaccidental serious harms of others up to individual choice, this suggests the preferability of a two-staged process whereby an initial PES with a default choice is used to spark acceptance of AVs, before eventually morphing into an MES once the widespread trust of AVs is secure and we are better able to understand how different ethical settings in AVs will play out in practice.

## Notes

1. By AVs we shall mean here, drawing on the SAE standards, Level 4 (when in autonomous mode) and Level 5 AVs only, as opposed to Level 2 or Level 3 systems that require humans to continuously monitor functioning and intervene in emergency cases such as those described here.

2. There are, however, difficult questions here in terms of how much risk to allow during testing (Wolkenstein 2018).
3. Indeed, it is so strong that it raises the question of whether humans will be *permitted* to drive once we have reliable AVs (Sparrow and Howard 2017).
4. The moral problems AVs face are a matter of *interaction* (they depend on the choices of others) and *iteration* (it is an ongoing policy), rather than the *one-off* choices we make that *only impact others* typically seen in trolley problems (Gogoll and Müller 2017). For more on trolley problems and AVs, see Himmelreich (2018), Hübner and White (2018), Keeling (2020), and Nyholm and Smids (2016).
5. This is sometimes claimed to be the *only* issue that an AV's "ethical knob" needs to address (Contissa, Lagioia, and Sartor 2017). But this ignores all the other ethical issues we outline here.
6. There is some preliminary evidence, based on a VR experiment, that people are willing to sacrifice themselves to save a group of five or more others, but not less (Faulhaber et al. 2019).
7. Previous studies have shown people have a preference to save children over adults (Faulhaber et al. 2019; Sütfeld et al. 2017).
8. On the first two views, a random number generator would have to be used as part of the decision-making process.
9. One might wonder how AVs could possibly do this. Perhaps, for example, they could use facial recognition software linked to various databases or use algorithms that estimate age and health based on appearance.
10. A third option, which we won't consider further, is that different manufacturers could offer AVs with different ethical settings that can't be changed by end users. This would offer limited personal choice, insofar as consumers could pick a different ethical setting package by buying from different manufacturers.
11. Note, we shall focus here only on "high-stakes" ethical settings, such as who or what a car should crash into, rather than "low-stakes" ethical settings, such as the temperature on a car's climate control system (e.g., different temperature settings use slightly different amounts of energy which minutely impacts climate change) since the later decisions are arguably best left to individual control (Millar 2017).
12. A study by Bonnefon, Shariff, and Rahwan (2016, p. 1573) showed that, while people approve of "utilitarian" AVs that sacrifice their passengers for the greater good, they "would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs." Further, "participants disapprove of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV. Accordingly, regulating for utilitarian algorithms may paradoxically increase casualties by postponing the adoption of a safer technology."
13. Clearly empirical research is needed to test this claim.
14. One might worry either that this makes the standard for fully informed consent too hard to meet (which might be a problem in other areas, such as patient consent in medical contexts) or that this overemphasizes the importance of fully informed consent at the expense of other relevant issues such as trust. But both worries at best minimize the strength rather than negate the point of the complexity argument, as they

concede the point behind the complexity argument, that this is too complex a choice to be fully informed about, but contest how much of a problem this is in practice.

15. As Wolkenstein (2018) writes: “We do not have enough knowledge, nor adequate decision rules, to decide about ethics in advance, without compromising the benefits of technological progress.”

## References

Bonnefon, J.-F., Shariff, A., and Rahwan, I. 2016. “The Social Dilemma of Autonomous Vehicles.” *Science* 352, no. 6293: 1573–76.

Contissa, G., Lagioia, F., and Sartor, G. 2017. “The Ethical Knob.” *Artificial Intelligence and Law* 25, no. 3: 365–78.

Faulhaber, A. K., et Al. 2019. “Human Decisions in Moral Dilemmas Are Largely Described by Utilitarianism.” *Science and Engineering Ethics* 25, no. 2: 399–418.

Formosa, P. 2017. *Kantian Ethics, Dignity and Perfection*. Cambridge: Cambridge University Press.

Formosa, P., and Ryan, M. 2021. “Making Moral Machines: Why we Need Artificial Moral Agents.” *AI & Society* 36: 839–51.

Gerdes, J. C., and Thornton, S. M. 2016. “Implementable Ethics for Autonomous Vehicles.” In *Autonomous Driving*, ed. M. Maurer et al., 87–102. Berlin: Springer.

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

Hevelke, A., and Nida-Rümelin, J. 2015. “Responsibility for Crashes of Autonomous Vehicles.” *Science and Engineering Ethics* 21, no. 3: 619–30.

Himmelreich, J. 2018. “Never Mind the Trolley.” *Ethical Theory and Moral Practice* 21, no. 3: 669–84.

Hübner, D., and White, L. 2018. “Crash Algorithms for Autonomous Cars.” *Ethical Theory and Moral Practice* 21, no. 3: 685–98.

Jenkins, R. 2016. *Autonomous Vehicles Ethics & Law*. *New America Foundation*. https://www.newamerica.org/digital-industries-initiative/policy-papers/autonomous-vehicles-ethics-law/.

Keeling, G. 2020. “Why Trolley Problems Matter for the Ethics of Automated Vehicles.” *Science and Engineering Ethics* 26: 293–307.

Kerstein, S. 2013. *How to Treat Persons*. New York: Oxford University Press.

Lin, P. 2016. “Why Ethics Matters for Autonomous Cars.” In *Autonomous Driving*, ed. M. Maurer et al., 69–85. Berlin: Springer.

Millar, J. 2014. “You Should Have a Say in Your Robot Car’s Code of Ethics.” *Wired*. https://www.wired.com/2014/09/set-the-ethics-robot-car/.

Millar, J. 2015. “Technology as Moral Proxy.” *IEEE Technology and Society Magazine* 34, no. 2: 47–55.

Millar, J. 2017. “Ethics Settings for Autonomous Vehicles.” In *Robot Ethics 2.0*, edited by P. Lin, K. Abney, and R. Jenkins, 20–34. New York: Oxford University Press.

Nissenbaum, H. 2011. “A Contextual Approach to Privacy Online.” *Daedalus* 140, no. 4: 32–48.

Nyholm, S. 2018. “The Ethics of Crashes with Self-Driving Cars.” *Philosophy Compass* 13: e12507.

Nyholm, S., and Smids, J. 2016. “The Ethics of Accident-Algorithms for Self-Driving Cars.” *Ethical Theory and Moral Practice* 19, no. 5: 1275–89.

Santoni de Sio, F. 2017. “Killing by Autonomous Vehicles and the Legal Doctrine of Necessity.” *Ethical Theory and Moral Practice* 20, no. 2: 411–29.

Scheutz, M. 2016. “The Need for Moral Competency in Autonomous Agent Architectures.” In *Fundamental Issues of Artificial Intelligence*, edited by V. C. Muller, 517–27. Springer. https://doi.org/10.1007/978-3-319-26485-1_30

Smith, P. T. 2019. “Just Research into Killer Robots.” *Ethics and Information Technology* 21: 281–93.

Sparrow, R., and Howard, M. 2017. “When Human Beings Are Like Drunk Robots.” *Transportation Research Part C* 80: 206–15.

Sütfeld, L. R., et al. 2017. “Using Virtual Reality to Assess Ethical Decisions in Road Traffic Scenarios.” *Frontiers in Behavioral Neuroscience* 11: article 122.

Thaler, R. H., Sunstein, C., and Balz, J. P. 2012. “Choice Architecture.” In *The Behavioral Foundations of Public Policy*, edited by E. Shafir, 428–39. Princeton, NJ: Princeton University Press.

Turner, P. N. 2014. “ ‘Harm’ and Mill’s Harm Principle.” *Ethics* 124, no. 2: 299–326.

van Wynsberghe, A., and Robbins, S. 2018. “Critiquing the Reasons for Making Artificial Moral Agents.” *Science and Engineering Ethics* 25: 719–35.

Wolkenstein, A. 2018. “What Has the Trolley Dilemma Ever Done for Us (and What Will It Do in the Future)? *Ethics and Information Technology* 20, no. 3: 163–73.

World Health Organization. 2018. *Global Status Report on Road Safety 2018*. https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.