NEW AMERICA

RYAN JENKINS

AUTONOMOUS VEHICLES ETHICS & LAW

Toward an Overlapping Consensus

SEPTEMBER 2016

About the Authors

Ryan Jenkins studies the moral dimensions of technologies with the potential to profoundly impact human life. He is an assistant professor of philosophy and a senior fellow at the Ethics & Emerging Sciences Group at California Polytechnic State University, San Luis Obispo. His interests include driverless cars, algorithms, autonomous weapons, and military ethics more broadly. His work has appeared publicly in *Forbes*, *Slate* and elsewhere, and he is currently co-editing two books on military ethics and robot ethics, both for Oxford University Press. At Cal Poly, Jenkins teaches courses in ethics, political philosophy, and the philosophy of technology, among others. He earned his B.A. in philosophy from Florida State University, Phi Beta Kappa, and his Ph.D. in philosophy from the University of Colorado Boulder.

Acknowledgments

A special thanks is due to Colin McCormick and New America fellow Levi Tillemann for comments on an earlier draft.

Cover image: DimiTVP/Wikimedia.

About New America

New America is committed to renewing American politics, prosperity, and purpose in the Digital Age. We generate big ideas, bridge the gap between technology and policy, and curate broad public conversation. We combine the best of a policy research institute, technology laboratory, public forum, media platform, and a venture capital fund for ideas. We are a distinctive community of thinkers, writers, researchers, technologists, and community activists who believe deeply in the possibility of American renewal.

Find out more at **newamerica.org/our-story**.

About the Digital Industries Initiative

The Digital Industries Initiative of New America brings together leading experts and policymakers from the private sector, government, universities, and other nonprofit institutions and the media, to analyze and debate the future of America's major economic sectors in the Digital Age. Each month the invitation-only Digital Industries Roundtable, hosted at New America's headquarters in Washington, D.C., features a discussion of the challenges of innovation in a different American industry. In addition, the Digital Industries Initiative publishes groundbreaking reports, hosts public events, and undertakes other activities at the dynamic intersection of technology, economics, and public policy.

Contents

Abstract	2
Introduction	3
Crash Optimization	4
Overlapping Consensus: Ethics as an Engineering Problem	6
Proposals for Crash Optimization	10
Adjustable Ethics Settings	14
Insuring Autonomous Vehicles	16
Handing off to the Driver	17
Abuse	19
Far-term Issues	21
Next Steps	22
Works Cited	24
Notes	26

ABSTRACT

There is a clear presumptive case for the adoption of autonomous vehicles (AV). It is widely believed they will be safer than human-driven vehicles, better able to detect and avoid hazards and collisions with other drivers and pedestrians. However, it would be unreasonable to expect AV to be *perfect*. And unlike programming for much software and hardware, the conditions AV can be expected to face on the road are an "open set": we cannot exhaustively test every scenario, since we cannot predict every possible scenario. In light of this, we must think carefully about what requirements manufacturers should have to demonstrate before AV are allowed on the roads. This paper surveys the practical state of the art, technical limitations of AV, the problem of driver handoff, and the possibility of abuse with AV, such as other drivers playing "chicken" with AV. It considers AV from the legal, ethical, and manufacturing perspectives before arguing for an "overlapping consensus": AV that behave in ways that are morally justified, legally defensible, and technically possible. The paper closes by applying this lens to some possible ways that AV could behave in the event of a crash, offering tentative endorsements of some of these, and recommending a closer collaboration between industry and the academy.

This report was inspired by the Autonomous Vehicles & Ethics Workshop held at Stanford University in Palo Alto, California in September of 2015. The workshop was a closed, invitation-only meeting of about 30 participants. Participants included academics, including ethicists, psychologists, roboticists and mechanical engineers; insurance lawyers and legal experts; and representatives from the automotive industry and Silicon Valley. The conference was organized by Patrick Lin (California Polytechnic State University), Selina Pan (Stanford), and Chris Gerdes (Stanford), and was supported by funding from the US National Science Foundation, under award no. 1522240. The meeting was conducted under The Chatham House Rule, whereby participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s) may be revealed without their expressed consent. This report includes input and observations from the workshop's participants. Its interpretations of those remarks, substantive claims and recommendations, however, solely reflect the syntheses of its author, and do not necessarily reflect the views of the workshop's participants, organizers, or supporting organizations. A special thanks is due to Colin McCormick and New America fellow Levi Tillemann for comments on an earlier draft.

INTRODUCTION

There is a clear presumptive case for the adoption of autonomous vehicles (AV). It is widely believed they will be safer than vehicles driven by humans. For example, AV will not become sleepy, distracted, or angry behind the wheel, and will be better able to detect and avoid hazards and collisions with other drivers and pedestrians. Because car accidents kill around 30,000–35,000 people per year in the United States alone[1], and because around 94% of crashes are due to driver error[2], the case for AV from increased safety and lives saved is extremely compelling.

Even mildly optimistic predictions concerning AV show that they could provide significant benefits in terms of social costs of death or injury, as well as increased convenience and productivity time for the individual consumer.

However, it would be unreasonable to expect AV to be perfect. Software and hardware undergo continuous development, and their failures are sometimes catastrophic. Since there is nothing intrinsically different about the software and hardware to be used in AV, the same possibility for catastrophic failure exists—witness the failure of Tesla's autopilot system in May, 2016 (Tesla Motors, 2016). And unlike programming for much software and hardware, the set of conditions AV can be expected to face on the road is an "open set": manufacturers cannot exhaustively *test* every scenario, since they cannot *predict* every possible scenario. Manufacturers will be unable to ensure that AV are totally prepared to drive on their own, in all conditions and situations.

In light of this, stakeholders must think carefully about what requirements should be met before AV are allowed on the roads. What kind of discrimination capabilities should AV have before it's permissible to deploy them? Is it merely enough that AV be superior to human drivers? How should AV be programmed to behave in the event of a crash, and is it permissible for them to change the outcome of a crash by redirecting harm? Or should we be worried about people who are killed or injured by AV when they would not have been otherwise? These and other issues are explored below, synthesizing the perspectives of philosophers, lawyers, and manufacturers, in search of an overlapping consensus on the development and deployment of autonomous vehicles.

CRASH OPTIMIZATION

Sensing the World

Autonomous vehicles use a variety of technologies to detect the outside world, make sense of it, and make decisions based on that data. Manufacturers like Google and Tesla take different approaches to which sensors they use, how data from those sensors is synthesized, and how AV make decisions based on the data. A typical suite of technologies for sensing the world includes cameras, ultrasonic sensors (SONAR), RADAR, or LIDAR (light detection and ranging). For example, Tesla prefers to use a combination of cameras and RADAR over Google's LIDAR because LIDAR instruments are orders of magnitude more expensive, even though they offer a higher resolution representation of an AV's surroundings.[3] Moreover, cameras are vulnerable to some of the same failings as the human eye: they have difficulty seeing in direct sunlight, low-light conditions, or inclement weather.

AV take in this information and synthesize it into a picture of the world through computer vision technology, recognizing lane markings, other cars, and obstacles on the road (Wernle, 2015). Finally, AV learn to navigate through a combination of "top-down" instruction and "bottom-up" machine learning. AV may learn, for example, by watching footage of human drivers, synchronized with data about pedal and steering wheel inputs, to mimic human behavior (Shapiro, 2016). Once implemented, behavior is continually reinforced or modified by subtle human nudges (Fehrenbacher, 2015), and can be fine-tuned with hard-coded inputs, for example, about the distance to keep between cars, the position to maintain in a lane, how long to wait after a stoplight turns green, and so on. With this suite of technologies, AV have already been shown to be able to reliably navigate roads in some situations.

"No-win" Scenarios

Autonomous vehicles presumptively bring significant benefits: their increased computational power and reaction time allow them to avoid some accidents that a human driver could not. A clear primary goal for AV should be to avoid all collisions. If AV could successfully avoid most or all accidents that are caused by driver error, this could be expected to eliminate tens of thousands of deaths and injuries each year in the United States alone.

However, it is possible that there could also be so-called "no-win" situations: situations where a crash is inevitable. Imagine, for example, that an AV is driving on the highway, and is boxed in on either side, with a car bearing down on it from behind as well. Should the vehicle in front of it slam on the brakes, the AV could have nowhere to go to avoid a collision. In these cases, an AV may

have the opportunity to "optimize" the crash by aiming for some goal. For example, one initially plausible overriding goal for AV is to minimize harm to humans. However, choices about how to direct or distribute harm are significantly morally freighted and demand extraordinary scrutiny.

If an AV cannot avoid a crash, then perhaps it should prioritize the protection of certain kinds of targets over others. A hierarchy of moral importance among potential targets naturally suggests itself: above all, avoid colliding with pedestrians or unshielded people, then avoid people on bikes, then avoid people in cars. This hierarchy is framed to do the most to protect the most vulnerable people on the road.

Inanimate Objects and Economic Damage

Inanimate objects would be given the least importance in a system that ranks harms in this way. But there are complications even with telling animate from inanimate objects: perhaps steering into what appears to be an inanimate object entails steering into a baby stroller or a propane tank on the side of the road. Proposals for crash optimization face significant technical challenges in being able to distinguish between different kinds of objects on the road.

Even supposing that the AV can correctly distinguish between animate and inanimate objects, and direct itself only toward inanimate objects, there are further questions about how inanimate objects should be treated in such a ranking. Suppose an AV is returning to its home, empty, while dropping off its owner at work, and suppose it is faced with a decision to steer off the side of the road, doing significant damage to itself, or colliding with another, occupied car, doing less damage to both vehicles. Should the AV should assign *itself* any weight in the decision? The idea that an empty AV should willfully sacrifice itself in the event of a crash should give us pause, though in this case it would minimize the risk of harm.

The economic benefits of avoiding crashes should not be overlooked, either, since crashes cause hundreds of billions of dollars in damage each year.[4] Would it be legitimate, then, to direct harm toward less expensive cars, or older cars, to reduce the total amount of economic damage resulting from a crash? Should an AV avoid hitting another car from the same manufacturer to protect the economic interests of its own manufacturer? Many professional engineering societies restrict the reasons on which an engineer can discriminate in their designs. Should reasons like this be struck from ethical programming?

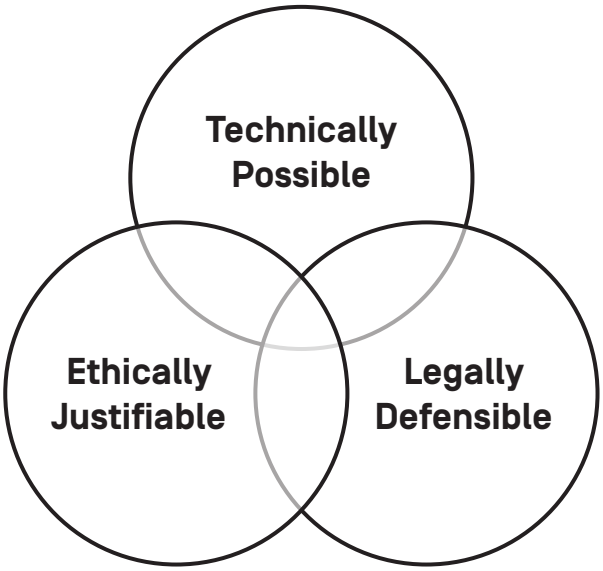
Problems like this provide a reason for external regulation, or at least a great degree of transparency, about which classes of objects are taken into account in automated ethical decisions. These cases show that some ways of distributing harm, even among inanimate objects, may be seen as unjust. More exploration by ethicists is required here to determine the legitimate bases on which AV could optimize a crash.

OVERLAPPING CONSENSUS: ETHICS AS AN ENGINEERING PROBLEM

The design of automated ethical decisions involves the interests of many different stakeholders: consumers, manufacturers, programmers, industry and executives. Each of these stakeholders will have different sets of motivations and constraints. It may be impossible to devise a "single best" algorithm that totally satisfies the preferences of these different groups, especially given that their preferences can conflict. For example, designing a car that is maximally safe, while preferable for consumers, might be prohibitively expensive or unprofitable, and thus unacceptable to manufacturers.

Rather than searching for a single ethical principle that should guide the programming of AV—a so-called "prime directive"— programming ethics should be considered as an engineering problem that allows a range of solutions. This problem requires *optimizing along several dimensions while obeying some strict constraints*. Constraints rule certain options out, for example, because they are practically impossible or morally unacceptable. By identifying constraints and common interests, stakeholders could hope to arrive at an overlapping consensus, providing a range of acceptable positions.[5] Limits on programming, the current status of the law, and ethical complexity restrict the possible options. But preserving a range of possible options within the overlap respects consumer autonomy and allows for competition between manufacturers.

To determine which decisions should be left to consumers and which should be "hard-coded" by engineers, decisions can be divided into high stakes and low stakes. High stakes decisions are more morally important than low stakes decisions, and so require more care on the part of engineers, and input from a greater range of actors. An example of



a high stakes moral decision would be how an AV ought to behave in a crash, including whether it ought to steer itself so that the resulting harms are distributed in a certain way. Low stakes decisions are less morally important, and there is less concern if they are hard coded by engineers and taken out of the hands of consumers. An example of a low stakes decision is the question of who should go first at a four-way stop. This decision may seem trivial, but that is the point: it could certainly have some morally important consequences, since it could determine who gets to work on time and who is late. Still, it is clearly less morally important than other questions at issue. This discussion will focus on high-stakes decisions. Crash optimization decisions are the clearest cases of high stakes automated ethical decision making.

Practical Limitations

The current state of sensor technology and the limited ability to represent complex decisions in computer programming constitute practical limitations on the ethical programming that is available and what we could reasonably demand from an AV. Current computer vision technologies cannot distinguish the objects that humans can distinguish as reliably as humans can: for example, AV might not be able to distinguish a speed bump from a person lying in the road, even though it would surely be *morally better* if it could. This is a practical limitation to ethical programming that is grounded in the limitations of the technical state of the art.

Technological Limitations

People usually speak as if *driving* is a single activity, and ask whether AV will be able to "do it" as well as humans or not, as if the answer were binary. However, this sloppy language obscures the full variety of tasks involved in driving and the range of contexts in which people drive. Driving requires apprehending the surrounding area, making inferences about the behavior of other humans and other objects, and an advanced (if implicit or subconscious) understanding of physics. Compare, for example, the difference between driving on a deserted highway with a clear median and lane markings at a constant speed and in bright daylight, versus navigating a construction zone with no lane markings in the middle of a blizzard on frenetic downtown Chicago streets.

Driving also requires making inferences about human intentions, for example, when a driver encounters a construction zone with a worker directing traffic and is instructed to drive on the wrong side of the road. The development and adoption of AV requires, then, not only that humans trust AV, but that AV can understand and trust the directions of humans when appropriate.[6]

There are significant and storied technical difficulties currently haunting manufacturers. Current sensors are lacking in important ways: for example, AV are unable to operate reliably in inclement weather or direct sunlight. And if an AV is going to brake for every paper bag or balloon in the road then it is not a usable system.

One solution is that AV should have the ability to judge for itself whether conditions are within its safe operating parameters and then refuse to initiate autonomous driving. An AV may refuse to drive, however, if it is raining or snowing heavily. But this safety check is of no use if weather or conditions or lane markings should change suddenly after the AV is already operating autonomously. Human psychology makes "handing off" to the driver challenging (this is explored more below) and, at any rate, if an AV is unoccupied, this could result it in being stranded.

The Limits of Ethical Programming

When choosing to program certain moral algorithms into AV, programmers must be able to represent that moral theory as a list of discrete instructions or mathematical values. In order to program a machine to obey moral rules, those rules must be *formalizable*. Some philosophers doubt that morality can be formalized in a list of

discrete rules that exhaustively cover all situations a person (or AV) might encounter (McDowell, 1979). What this means is that AV might have to operate according to impoverished, abbreviated, or oversimplified moral rules, which are much less powerful, informative, and nuanced than those a human driver would use.[7] However, efforts have been underway since the middle of the 20th century to provide mathematical representations of moral decision-making. These techniques have been lying in wait for the moment they could be put to use, i.e. by embedding them in machines.

We need not contemplate a far-future where AV can recognize and classify objects as well as humans can, and can navigate morality as humans can: manufacturers are hindered in even approaching this future by the current state of technical sensors and roadblocks in ethical programming. Both of these limitations are non-negotiable in a sense. Limitations of the technological state of the art are not open to debate, in the way that laws may emerge from discussions among interest groups.[8] Moreover, if the best moral theory cannot be captured by a simple list of rules, that is a necessary feature of morality and computing, not a technical problem waiting to be solved (Purves, Jenkins, & Strawser, 2015). These are the most stringent, unavoidable constraints on programming the behavior of AV.

While it may be impossible to describe the best moral theory as a series of discrete rules, this is not the only (or the most likely) approach that engineers are likely to take. Companies are already using machine learning techniques to educate AV in how to drive. But machine learning threatens to lead to code that is inscrutably complicated, which creates new problems. Code that is overly complex can become difficult to interrogate deeply, maintain, debug. Confusingly complex "spaghetti-like" code has already been blamed in some suits against manufacturers.[9]

Policymakers may be prudent to place constraints on the complexity of code that determines the ethical programming of AV. Alternatively, industry professionals should develop best practices—and accountability—for ensuring their code is neither opaque nor inscrutable. Clearly, it would be difficult to enforce code legibility and transparency through a government body composed of bureaucrats or other experts. Still, a company that produces a product that confuses its own engineers should be held liable for the unpredictable or unreliable functioning of its product. Companies should take pains to make sure their code is transparent, for example, in something like DARPA's explainable artificial intelligence (XAI) (Gunning, n.d.) or CMU's quantitative input influence (QII) measures (Spice, 2016). Ethical programming that is beyond the understanding of even teams of experts should be prohibited or disincentivized.

There is much at stake by recognizing these technical limitations. Some of the practical shortcomings of AV could have unfortunate consequences from the point of view of justice. While AV have been hailed for their ability to augment the capabilities of the disabled or blind, there are also worries that some of the most vulnerable people in society could be harmed by technical failures of AV or the oversight of manufacturers. For example, a child who wanders into the street, or a disabled person who has fallen into the street and can't get up, may be a victim of an AV's inadequate programming and sensor technologies. A blind passenger of an AV would be unable to take control of the AV if need be, and so could become stranded. Questions of programming and design become especially important here because they could exacerbate differentials in power and capability in society that many are continually trying to eliminate. This underscores the care necessary in designing inclusively, with deliberate attention paid to the effects of design decisions on historically marginalized or disadvantaged populations.

Legally Defensible Options

Manufacturers can be expected to support certain ethical programming only if they believe it would be legally defensible. It is important to note here

that the law should be expected to change and adapt to AV—part of the impetus of this report. Manufacturers will likely be agents of this change, lobbying for regimes that benefit them or insulate them from liability. Still, until this occurs, there is some uncertainty about how current laws apply to AV and how these cases will be litigated. Until this ambiguity is clarified, manufacturers should be expected to proceed cautiously and defensively. While the law should be expected to change, manufacturers should also be expected to operate within its apparent strictures for the time being.[10]

The law typically errs on the side of *not doing harm*. This precedent might narrow the available choices for ethical programming considerably: namely, by ruling out any "crash optimization" program that intentionally steers toward another person or car, even if it minimizes the total harm that results. In this case, a judge would probably conclude that the AV made things worse by steering into the one person, specifically, made them worse *for that person*, and would hold the manufacturer liable for that death. Such an argument might go like this: "This car *steered into* someone, and so it *did harm*. The alternative, maintaining its course while slamming on the brakes, would have resulted in more harm, but at least the car would not have *aimed* at any one of the people injured."

Philosophers have long disputed this distinction between doing harm and merely allowing harm to occur through inaction, yet this reasoning relies on this controversial distinction (Foot, 1967; Thomson, 1976; Quinn, 1989). It may be doubted, for example, whether an AV that *sustains, maintains, or accepts* a course of action that leads to several deaths merely *allowed* those harms, or whether it *caused* them (Bennett, 1998; Norcross, 1999). Is it really accurate to say that an algorithm merely allows something to happen if it *anticipates the outcome* of its current course, considers changing its course, and then does not? Some philosophers have argued that people can also accomplish things through *inaction*. For example: a doctor may kill her patient by *withholding* treatment from him, or the Secretary of State may offend a foreign dignitary by *not* shaking her hand.[11] If merely allowing harm is adjudicated legal, and if this leads to outcomes that are significantly harmful, perhaps the law should change to allow AV to *actively direct* their harm instead when it would save overall lives.

From the legal point of view, this point may ultimately be moot: if an AV kills one person, the manufacturer will get sued; if it instead *redirects* itself toward one person, the manufacturer will simply face a different plaintiff. It may be most important to identify which kinds of ethical programming could survive cross-examination in front of a jury. Suppose an investigation is launched into the programming of an AV involved in a crash. In order to placate jurors, manufacturers will have to bring evidence of all the accidents that were *avoided* by their particular choice of ethical programming—to argue, as it were, that the *overall benefits* of the design choice justify the occasional accident, injury, or death. There is need for empirical research on this question, for example, collecting data on jurors' opinions of fault in analogous cases, or by conducting controlled experiments in a simulator. Such research could ground future legal rulings about how a reasonable person would ascribe moral responsibility in such cases.

One interesting finding from early investigations is that, when confronted with an autonomous driving accident, observers tend to blame actors *not* at the scene. The "bubble" for responsibility would typically be drawn around people who are at the scene, such as one (or both) of the drivers of the cars involved in the accident. When considering AV, the bubble moves to others who are not present, such as programmers, manufacturers, and engineers, or even abstract entities such as government regulatory agencies. Philosophers have considered such questions of "collective responsibility" for some time (Feinberg, Collective responsibility, 1968). The law may soon need to examine these views in order to fairly adjudicate cases like this, and to make sure these cases are settled in a way that is justifiable to both the public and manufacturers.

Ethically Acceptable Options

Reaching a consensus regarding the ethical programming of AV is a supremely challenging task. For the foreseeable future, the only workable solution is to allow engineers and manufacturers some leeway to stake out a position within a range of ethically acceptable options. Philosophers' longstanding conversations on the justice of distributing benefits and harms will be invaluable for delineating this ethically acceptable sphere (Walzer, 2008; Roemer, 1998).

Allowing a great deal of variety of ethical decision procedures respects both manufacturer and consumer autonomy, and secures the benefits of market competition among firms. But not just any ethical decision making algorithm is morally permissible, and so the range open to manufacturers should be constrained by moral considerations. As it is today, the freedom of corporations should be restricted to honor deeply held non-monetary social values such as the value of human life, fairness, and safety.

PROPOSALS FOR CRASH OPTIMIZATION

Autonomous vehicles may be forced to distribute harms among people in "no-win" scenarios, for example, by deciding whom should be harmed in a crash. Thought experiments can help us evaluate the acceptability of various methods of distributing harms and benefits. These thought experiments are often called "trolley problems." These are named after a famous case in ethics—which has generated a veritable cottage industry among academic moral philosophy—that imagines a runaway trolley careening down a track toward five people, with the option to switch the track to one person. Should you switch the track to save the five people, in the process killing the one person (Foot, 1967; Thomson, 1976)? Variations of this case have rapidly multiplied since it was originally introduced in the 1960s, and are useful for "stress-testing" ethical principles. If the proposed ethical principles generate unacceptable implications in imaginary cases, then those principles should not be embedded in AV in the real world.

Inaction

Suppose an AV is careening toward five people, but can steer itself toward one. Steering itself toward one could count as *killing* the one person, whereas

hitting the five would only count as *letting them die* through inaction. Inaction like this may be the only legally defensible option, as discussed above, since the law has traditionally erred on the side of not inflicting harm.

Moral philosophers have vigorously contested the distinction between performing an action and merely allowing a result to come about (Rachels, 1975; Norcross, 1999). For example, there may be an argument that the AV does actively do something when it accepts, maintains, or sustains a path toward the original five people, and so it is not obvious that the AV's programming merely lets them die. A legal argument could be based on this moral one. There may be a legal duty of manufacturers to show "due care," which is flouted with regard to the five people, since the AV has the opportunity to respect their wellbeing but does not.

Even if there is an important moral difference between bringing about a harm versus allowing that harm to come to pass, most moral philosophers argue that there is a moral duty to direct our harm toward the smaller party, thereby minimizing the total amount of harm that comes about (Bourget & Chalmers, 2014). Intentionally killing one person is probably preferable to foreseeing that five people will die and then allowing it. If "inaction" means accepting these five deaths rather than intentionally bringing about one, then most ethicists would reject "inaction" as morally unacceptable.[12]

Harm Minimization

The proposal to minimize the total amount of harm that results from a crash follows naturally from the rejection of "inaction." However, this proposal also faces significant challenges. Note, for example, that there is an important difference between minimizing the *number of people who are harmed* versus minimizing the *total harms* that come about. If I seek to minimize the number of people who are harmed in a crash, then I might do that by killing one person rather than slightly injuring two others. It is probably a superior proposal to minimize the total harms that come about, no matter how those harms are distributed between people. However, implementing this is practically impossible for the time being, since AV cannot be expected to accurately anticipate how many people will be injured rather than killed in a crash, not to mention the severity of the injuries. Manufacturers would need much more data and exquisite physics calculation ability available to AV to make these kinds of predictions. Other data like the number of passengers potentially involved in a crash, whether they are pregnant, whether they are wearing seat belts, etc., would also be crucial. AV shouldn't be expected to be able to make these fine-grained distinctions for the near future.

Maximin

"Maximin" is a candidate definition of rationality: it says that for a person to choose rationally between several options, they should optimize their worst possible outcome, so that the option they choose has the least bad potential outcome of any of the options available (Rawls, 1974). The principle is loss-averse, and gives preference to avoiding losses over maximizing potential gains.

In the case of AV ethical programming, a maximin decision procedure would require an AV to behave to minimize the worst harm that comes about by the crash. Or: to behave so as to make the person who is injured most by the crash as well-off as possible. Imagine an impending collision between a car and an AV. Suppose if the AV maintains its course, some passengers could be killed and others would be spared; whereas if the AV swerves to one side, all passengers risk minor injury. A maximin procedure would require the AV to swerve since the worst possible injury that could result when swerving, i.e. minor injury, is less bad than the worst possible injury that could result when maintaining the course, i.e. death.

Maximin is initially plausible but faces some significant objections. For example, it might require that an AV take a path that will injure its own passengers, even when they wouldn't

otherwise be injured. Many consumers would balk at this possibility. Maximin fails to take account of legality, for example, and might swerve to avoid reckless drivers playing "chicken" with an AV, even though many would think that those drivers have made themselves liable to some harm by intentionally breaking the law. Finally, maximin could endorse behaviors that injure many dozens of people—imagine a school bus full of children—rather than subject a single person to a more serious harm. These objections point to the need for taking the legality of various actors into account and for aggregating the harms that could result from a crash, which are both considered in the next proposal.

Legality-adjusted aggregate harm minimization (LAHM)

Finally, let's consider a theory that improves on the previous theories and avoids their shortcomings. This theory considers both the total amount of harm that results from the crash as well as the legality of the various actors involved. *Legality-adjusted aggregate harm minimization (LAHM)* seeks to minimize the total harm that results from a crash, where the value of each person involved is sensitive to whether they are obeying the law.[13] For example, the passengers of a car that drive across a double-yellow line would be liable to more harm than the passengers of a car that is obeying all traffic laws.

LAHM avoids problems that maximin faces. It would not distribute harms in a way that leads to greater overall harm, for example, by preferring many small harms to many more people, unless doing so could be justified by appeal to the legality of the actions of the people affected. It would discount drivers who are playing chicken, acknowledging to popular intuitions that drivers *at fault* are somewhat responsible for their own harm or that, if someone must be harmed by a crash, it is *less regrettable* that they are harmed.

LAHM also provides a good moral reason why it would be permissible for an AV to harm its own driver or occupants. If the driver of the AV himself is breaking the law, then he is liable to be harmed.

LAHM enjoys intuitive support but faces many problems of its own. It was already noted above that, without fine-grained data about the injuries likely to result from a crash, these kinds of calculations are probably impossible. Philosophers would need a reliable way to assign weighted values to harms, for example, by assigning values to minor versus major injuries, and asking how injuries it would take to equal the badness of one death. They would also need to fix the "discount" value of someone breaking the law: is someone more egregiously breaking the law—say, by dangerously speeding—to be penalized to a greater extent?

Legality-weighting might also conflict with our broader motivation to minimize harm. LAHM could require aiming at a motorcyclist who is wearing a helmet rather than one who is not, since that would minimize the total resulting harm. But this means intentionally targeting the person who is obeying the law. This is a significant tension within the theory that will need to be resolved.

For an AV to evaluate the legality of the various behaviors involved, it would need to operate as judge and jury in a sense, applying byzantine traffic laws. It is unlikely in the near future that AV would be able to judge *mens rea*, i.e. a guilty intention, on the part of drivers, so they would be restricted to judging the legality of a car's behavior by some objective standard, or a third-party point of view. There is a significant legal distinction between violating the law through malice versus ignorance, but this will be beyond the ken of AV.[14] For example, there is a significant difference between speeding down a one-way street out of confusion and out of an intention to injure innocent people. LAHM could thus result in unjustified harms to new drivers or the elderly. And even judging the objective legality of other drivers' behavior would be difficult in some cases. For example, medical doctors are permitted to speed in California, but an AV could not reliably judge whether a speeding car is being driven by a doctor.

Strict Equality

These difficulties call legality-weighting into question. For the time being, then, AV should treat all drivers and potential crash victims as innocent and equally morally important. Until AV can reliably distinguish how many people actually occupy all of the cars involved in a crash, they should weight all cars equally, using the cars themselves as proxies for potential accident victims.

As perhaps the most difficult question facing AV manufacturers, various stakeholders will have to consider which ethical programs are acceptable to program into an AV. Every proposal—including simple inaction—is likely to generate some counterintuitive verdicts. For the near future stakeholders should aim to narrow the domain of possible decision procedures in order to define the sphere of technically feasible, philosophically-informed, and legally defensible options for crash optimization.

The Shifting Consensus

In closing this discussion about the search for overlapping consensus, it would be wise to acknowledge that the constraints and preferences that help us navigate these overlapping domains are sure to shift with the adoption, penetration, and incremental improvement of AV. The sphere of technical capabilities will certainly change, as programming techniques and sensor technologies advance. The sphere of legal acceptability may change as juries become more comfortable with autonomous driving, or as their focus shifts from individual harms to society-wide benefits. Alternatively, the sphere of legally defensible choices may constrict as the technical capabilities of AV advance: as more sophisticated programming and moral algorithms become possible, the public may demand more from manufacturers and be less forgiving of mistakes. The sphere of morally acceptable options will change with advancing technology as well—since manufacturers can only be obligated to do what they are physically capable of doing.

ADJUSTABLE ETHICS SETTINGS

Suppose discussions about the ethical programming of AV reach an impasse, where the relevant stakeholders can agree on a range of possible options but no single option as best. It has been suggested that drivers should be allowed to *select* an ethical program from a menu of acceptable options. Commentators have proposed allowing drivers to choose, for example, how generously or selfishly their AV should perform in the event of a crash. Should their AV give ultimate preference to protecting its passengers, should it distribute harm more equally, should it seek to protect the most vulnerable, etc.?

Many arguments point in favor of adjustable ethics settings. Ethical decision-making is typically an activity reserved for individuals, especially in the traditional driving context. It is a stark moral situation for a driver to place her hands on the wheel, and the car's behavior can be traced directly to her decision (or negligence). Humans are sensitive to being ordered around by machines—or at least to having their freedom restricted—for example, by being locked out of control. It is only natural for passengers sitting in the front seat of an AV (especially with a steering wheel in front of them) to feel responsible for avoiding obstacles in the road, and to feel disoriented if they are prevented from doing so. For consumers, knowing that they have some control over their AV's ethical behavior could be as important as the car's mileage.

The options for ethical programming discussed in the previous section shift ethical responsibility away from the individual toward other stakeholders not at the scene. Still, the driver may have preferences about how their car distributes benefits and burdens, and to codify a single ethical code would be to reduce the driver's autonomy. Providing adjustable ethics settings respects driver autonomy by devolving this decision-making back to its traditional source, the driver, along with the attendant moral and legal responsibility.

Some worry that drivers will not trust AV to make ethically important decisions on their behalf. This is significant, since high consumer trust is necessary for widespread adoption, which is in turn required to secure the benefits that AV promise. However, it is surprising how quickly people can come to trust a machine—the greater concern is consumers trusting them too much, as appears to have been the case with the driver killed in his Tesla Model S in May of 2016, after its Autopilot failed to brake.

Setting a "Moral Floor" for AV Behavior

A compromise, which restricts the freedom of drivers but honors their autonomy within an acceptable sphere is the most plausible position to adopt.

One suggestion is to follow a distinction made within moral philosophy as far back as Thomas Aquinas, and that is the distinction between actions that are morally required and actions that are morally good but not morally required, or actions that "go above and beyond" moral duty. These latter actions are called supererogatory (the word's origin is Latin, meaning *to pay out over and above* what one owes) (Feinberg, 1961). Most people agree, for example, that we are morally required to donate some of our money to charity in order to help those who are less fortunate. Most also agree that it is morally good, but *not* morally required, to donate *most* of our income to charity. This is an example of supererogatory self-sacrifice (Arthur, 1981).

Typically, people are not free to shirk their moral obligations, but they are free to not perform supererogatory actions. Applying the same thinking to adjustable ethics settings: in a crash, an AV must distribute the harms among possible victims in a way that is just, perhaps following one of the proposals discussed above. This requirement constitutes a "moral floor" below which AV are not allowed to go. Just as we should do our best to prevent human drivers from behaving in ways that are unjust or unfair (if we could), we should do the same in the case of AV.

Above that floor, drivers should have the freedom to adjust the ethical programming of their AV so that it behaves more sacrificially, for example, by taking on a greater amount of risk to its own passengers and thereby sparing others. But drivers should not be free to modify their AV's ethics settings so that it behaves more selfishly, so selfishly that it would distribute the resulting harms in a way that is unjust.

Informed Consent

Whichever ethical settings are left open to consumers, manufacturers should be sure to secure consumers' informed consent about the possible behavior of their AV. However, there are difficulties in knowing what amounts to informed consent and in manufacturers' abilities to properly educate consumers.

Informed consent requires informing consumers how an AV may perform in the event of a crash, since the behavior of the AV has consequences for their own safety and wellbeing. It is important that manufacturers manage consumer expectations by making clear the capabilities of an AV and the range of its autonomy. Besides being required by basic business ethics, securing informed consent also insulates manufacturers from liability. If manufacturers supply inadequate information about the possible behavior of an AV, there could be grounds for complaints about false advertising or deceptive trade practices.

However, it is unclear what informed consent amounts to in this case. Manufacturers currently hide information from consumers about the behavior of their cars. For example, if you consult a car's owner's manual to learn about its airbags, the manual may simply say, "In the event of a forward collision, your airbag may deploy." If such a nebulous warning is sufficient for informed consent, even in a case where the driver's wellbeing could be affected, then similarly vague warnings may suffice in the case of an AV's ethics settings.

If the overriding "prime directive" of the AV is to minimize harm to the driver, then this can be assumed to be covered under hypothetical consent (Stark, 2000). However, if the car behaves in any way that is unpredictable and potentially unjustifiable to the driver, then the driver should be made aware of that possibility. For example, an AV's manual might simply say, "In the event of a crash, your car may attempt to minimize the resulting harms to yourself and others by steering, braking, or accelerating." This statement glosses entirely the complex moral reasoning and programming taking place under the surface. While it need not disclose its entire ethical program—and this could compromise trade secrets—consumers should be made aware that their car could assume control in an emergency, and be given some general picture of what this might entail.

It is also unclear to what extent consumers will fully appreciate what they are buying into when they purchase an AV with adjustable ethics settings. Without elaborate demonstration or education—which is unlikely to come at the point of sale—consumers may be left in the dark about the ethical behavior of their AV. Expecting dealers to explain to consumers how an AV's ethical programming might work, or how its adjustable ethics settings might work, would be quixotic. Adjustable ethics settings would introduce a new level of confusion for consumers, and consumers could blame accidents on such confusion. The best way of securing the consent of consumers and the broader public is by developing ethics programs through an open, collaborative, and democratic procedure.

INSURING AUTONOMOUS VEHICLES

Insurance companies may be a reliable source of information for deciding between options for ethical programming in AV. The insurance industry already boasts the competencies for pricing risk, which are often taken to reflect larger social values.

However, the case with AV is complex. First, if the risks associated with AV are correlated, then insuring them is less likely to be a viable business model. For example, individual deaths are not correlated. However, damage or deaths from natural disasters like floods often *are* correlated. Companies are often reluctant to insure against correlated risks because the possibility of paying out many premiums at once is daunting. Risks associated with AV may be correlated if a single defect affects all the AV from one manufacturer or, worse, all of the AV on the road. This could happen if some standard code is mandated which turns out to be faulty or vulnerable to malicious attack.

Second, actuarial data about AV, which is necessary for pricing risk accurately, is still forthcoming. It may be several years after the arrival of AV until this data is readily available. Currently, much of the data concerning AV reliability and safety is held by manufacturers, who will have to share the data with insurers in order to price risk accurately.

Finally, insurance companies would likely demand from manufacturers the ability to know whether the driver or the autonomous programming is "in control" of the AV at any given moment. This is necessary for determining whether the driver or manufacturer (or some other third party) is at fault for an accident. This may raise questions about driver privacy, and will at least require special features to be built into the car, such as tracking the driver's activity and participation in driving.[15]

HANDING OFF TO THE DRIVER

As noted above, driving is not a single skill, but a suite of skills that must be exercised in a variety of circumstances. Because of this, some are skeptical that there will ever be an AV with the robust capabilities necessary to drive in all conditions. If this is true, then manufacturers and programmers must design AV to hand control over to a human driver (and human drivers must be prepared to take control) at a moment's notice. This is a unique problem of human-computer interaction, which requires intense study by psychologists, engineers, and designers.

This is especially problematic given that people tend to "use up" the safety that they are given by assuming greater and greater risks. There is already evidence that consumers behave this way with regard to seatbelts and automatic braking. This is known as *risk compensation*, and could pose special problems for driver handoff. Imagine, for example, that drivers feel free to get into their AV drunk, assuming their car will be able to ferry them home safely. In this case, the driver cannot be trusted to resume control of an AV.

Engineers should determine how much notice is appropriate to give drivers before expecting them to resume control of the AV. Drivers require at least two seconds of notice to reliably take control of an AV.[16] But, somewhat surprisingly, it is possible to give drivers too much notice. Giving drivers, for example, twenty seconds of notice may result in them looking up, searching in vain for the hazard, becoming confused or distracted, and returning their attention to their previous task. The "sweet spot" for capturing a driver's attention while not allowing them to lapse back into distraction seems to be somewhere between two and five seconds.

The amount of notice that drivers require has clear implications for the mechanical design of the AV, for example, in terms of the range of its sensors. Suppose, for example, that an AV is approaching an unexpected obstacle, and suppose it judges its own capabilities to be inadequate to negotiate the obstacle. If the AV needs to give a driver five seconds of notice, and it is travelling at 65 mph, then the AV's forward sensors must reliably judge obstacles 500 feet in front of the car. Some designs aim to give drivers as much as 10–15 seconds' warning (Hammerschmidt, 2015). This may be too long, given a typical driver's psychology, to reliably assume control, and also places significant technical burdens on manufacturers to develop sensors that can anticipate obstacles over a thousand feet front of an AV. This may prove unworkably demanding.

There are further questions about the best way to prepare drivers to take control of an AV in case of an emergency. It may seem obvious that the best advice is to require drivers to be paying attention to the road in front of the AV, even when they are not

driving. However, tests have shown that this can actually lead passengers to become drowsy or fall asleep, even after just a few minutes. This may be an example of the brain's tendency to "zone out" when observing a monotonous task with little interaction or engagement. Some research has shown, in fact, that the best approach may be to ask drivers to distract themselves with movies, games, reading, or activities besides driving to keep their minds alert. The conclusion of this research is captured in the slogan: *distraction becomes engagement in autonomous vehicles* (Miller, et al., 2015). Paradoxically, maximizing driver engagement and readiness to take control may mean telling them to do something *besides* pay attention to the road.

Drivers should also be expected to suffer from "skill rot," that is to say, their driving skills will have degraded after long periods of not driving. Humans can take anywhere from 15 to 45 seconds after having assumed control of an AV before their performance (e.g. their steering proficiency) stabilizes.[17] This problem will presumably be exacerbated the longer a driver has gone without controlling the AV. If a driver is in the middle of a 2,000-mile cross-country road trip, for example, this problem could be significant. There is also an important difference between being asked to assume control in the middle of a relatively safe situation, for example, traveling at low speeds on a city street, and being asked to assume control because the AV anticipates an impending emergency or collision. Moreover, AV can be expected to ask drivers to resume control *precisely* in those situations that are chaotic, unpredictable, or confusing.

Adding another layer of complexity to this research, however, it was found that even drivers who are drowsy or asleep can reliably retake control of an AV in an emergency. Clearly, more research is necessary to determine the safest activity or disposition for passengers to assume while the AV is in control of the car. It is highly unlikely that NHTSA or other legislative bodies will accept the paradox that the safest thing for a passenger to do is to *not* pay attention to the road.

Pilot Handoff in Commercial Aviation

Manufacturers can learn from the significant experience gained from research and practices in pilot handoff in commercial aviation. There, human pilots share control with autonomous flight systems, and autonomous flight systems routinely hand off to humans for takeoff and landing.

Pilot handoff in commercial aviation is an important analogy for AV, but is limited. Commercial airline pilots are trained extensively on the aircraft they fly. They are drug tested and legally required to get a certain amount of sleep before flying. In many countries pilots are required to have a co-pilot in the cockpit. For these reasons, commercial airline pilots constitute a best case scenario study for the preparedness and competency of human pilots.

Moreover, autonomous control of commercial airliners is easier than it is for cars, since the sky is big and airplanes are small. Driving on a road is an order of magnitude more challenging than flying in a mostly empty sky, with a great degree of freedom in three dimensions.

Still, human error is a major component in nearly all commercial airline incidents. Pilots can suffer from "mode confusion," where they are unsure which flight controls they are in command of, versus those the computer is controlling. By way of analogy, imagine a driver who owns two cars: one autonomous or semi-autonomous and the other not. Drivers would become accustomed to driving these cars in different ways, and so they could become disoriented when switching back and forth between driving their two cars. This could presumably cause them to drive recklessly or, at least, not as cautiously as they might in an AV.

One policy suggests itself, and that is a new licensing procedure for autonomous vehicles. Governments should consider requiring drivers to become re-licensed to drive AV. Alternatively, once AV become commonplace, governments should consider including a driver handoff in the standard test for a driver's license to familiarize the driver with the procedure.

ABUSE

The issues discussed thus far assume good faith on the part of the human actors involved. But this is an idealistic assumption, and society should also prepare for the full range of abuse that can be expected: how might human drivers behave badly, or intentionally misuse AV?

Playing "Chicken"

For one, abusive drivers might play "chicken" with AV: if they know that a car in the oncoming lane is autonomous and will swerve to avoid a collision, they could purposefully drive at the AV as a way of causing trouble. In the worst imagined cases, drivers could play chicken on narrow roads or cliff-sides, where an AV swerving could mean serious injury or death for its passengers.

Of course, this behavior is predicated on the expectation that AV will in fact swerve (at all costs to their passengers) to avoid a collision. This problem could be anticipated and avoided, then, by allowing AV to collide with drivers playing chicken, or at least by making the behavior of AV less predictable. Making their behavior less predictable in cases of chicken could introduce enough uncertainty to abusive drivers to make playing chicken undesirably risky.

However, both of these solutions have clear negative impacts. Consumers would presumably prefer AV that behave predictably, and would certainly not want an AV that would collide with oncoming traffic, even if that behavior is a necessary part of a larger pattern that dissuades abusive drivers from playing chicken. The values of transparency, and the autonomy and safety of drivers, conflict with this proposed solution. More work, particularly in game theory, is needed to explore practical and defensible ways of safeguarding AV against this kind of abuse.

This hypothetical scenario points to a more general problem. Autonomous vehicles may be programmed based on assumptions or data about human drivers' current habits. But, as AV become more common, humans could modify their behavior in response. New norms of driving could emerge. There will be a continuous interplay between autonomous driving and human driving, even as the number of human drivers dwindles to zero. AV programming and human behavior will likely proceed in a process of parallel evolution that calls for continuous reexamination.

Hacking and Hacking Back

Connections to the Internet are becoming increasingly common in cars, especially through their infotainment centers, but also through other hardware such as connected ECUs. Following this

trend, future cars including AV should be expected to sport even greater connectedness. Moreover, some of the presumptive benefits of AV are secured through vehicle-to-vehicle (V2V) networks, or vehicle-to-infrastructure (V2I) networks. These benefits include reduced traffic congestion, dense caravans that take advantage of drafting, and collision warning and avoidance. Manufacturers will have to take extreme care when designing these systems: the porousness of the Internet connections in cars has already been demonstrated (Greenberg, 2015).

Autonomous vehicles should be prepared to deploy "defensive countermeasures" against adversaries: changes in behavior designed to protect the safety and reliable operation of the AV and its passengers. Disconnecting from the Internet is an obvious suggestion. This would not violate the law, but would sacrifice many of the expected benefits of highly connected cars. Moreover, this solution is not without its moral drawbacks: for example, if the passenger is on the way to the hospital because they are injured, then this becomes problematic. Malicious actors could initiate hacks against AV as a way of accomplishing a "denial of service" attack, if they know an AV will "brick" itself, or segregate itself from the Internet, in its own defense. Bricking an AV could be paired with some other real-world attack as a means of marooning a person in need of help.

Stakeholders will have to consider whether it is worth sacrificing the benefits of having highly connected AV in order to prevent criminals from hacking into them. This is an open question that might only be solved once society has concrete experience of the benefits of highly connected cars as well as a clearer appreciation of the attendant risks.[18]

Autonomous vehicles could also be used as a deterrent to some kinds of abuse or law-breaking. For example, they could monitor the behavior of other drivers or cars and alert the police to drivers who are speeding, littering, or driving erratically. (Imagine a scenario where a driver gets a ticket in the mail, having been tattled on by an AV.) However, philosophers and lawyers by and large have been wary of ubiquitous surveillance. A policy of crowd-sourced law enforcement, carried out by autonomous machines, could have significant deleterious consequences for interpersonal trust and could foster feelings of resentment or powerlessness, not to mention false positives.

FAR-TERM ISSUES

This discussion has been limited to near-term issues for programming and manufacturing AV. Projecting the effects of disruptive new technologies into the long term is a more challenging task, and confidence about such predictions should be tempered accordingly. Still, a host of issues present themselves when imagining a future of autonomous driving.

Society can expect a significant number of jobs in transportation to be replaced by AV—perhaps numbering into the millions of displaced workers. This includes long-haul trucking, inter-city delivery, and taxi and chauffeur services. Cheap and reliable AV could erode the individual ownership model of cars and reduce the social and financial costs of being carless. Cities could enjoy greater density of planning with less need for dedicated parking spaces. Autonomous vehicles could remake the rural landscape as well, reducing the psychological distance between city and country, or between neighboring cities.

Today's built environment is, in many cases, the consequence of racist or classist urban design policies. For example, much urban design is intentionally unfriendly to bikes, pedestrians, or public transit, and some of these designs are relics of a time when explicit prejudice was accepted (Kolitz, 2015). Autonomous vehicles could assist us in our responsibility to reverse these historic injustices. It remains to be seen how helpful AV may be in this project, and what kinds of unforeseen negative consequences might arise from even this well-intentioned undertaking.

Autonomous vehicles' presumptive ability to avoid crashes and reduce deaths from car accidents is one of their most attractive features. But some of the unintended negative consequences of saving these lives are already predictable. For example, most donor organs come from car crash victims. As a result, a "zero-fatality" future could mean fewer patients benefitting from life-saving organ transplants (Griffith, 2014). Many rural hospitals and trauma centers derive a substantial portion of their income from these operations. Autonomous vehicles could cut into these sources of income, threatening the financial stability of some hospitals. Ultimately, this could reduce rural areas' access to healthcare. This is just one example of how the widespread adoption of AV, even if undoubtedly beneficial on the whole, could result in surprising and troubling harms to certain populations.

NEXT STEPS

Below are recommended next steps for developing trustworthy, informed, and morally defensible policy for AV.

Honing the Overlapping Consensus

Our highest priority should be honing and sharpening the overlapping consensus between stakeholders. This involves open, honest, and diligent collaboration between lawyers, programmers, manufacturers, technologists, and ethicists.

This process demands further discussions and workshops, including the consideration of specific, concrete decision procedures, informed by the technological state of the art. It may involve collecting qualitative data from various publics, such as consumers, lawyers, and ethicists, to gauge their support for or acceptance of decision procedures and approaches. It may demand psychological research on human passengers who are in simulated accidents involving AV, or mock trials involving people serving as jurors. This becomes increasingly important as the technical capabilities of AV mature—much of the ethical discussion above was hobbled by lack of fine-grained data about crashes or exquisitely sensitive AV sensors. In ten or twenty years, the suggestions above bracketed as unrealistic may be possible.

The parties involved must come to an agreement of what kinds of decision procedures for AV are technically feasible, which are legally defensible, and which are morally acceptable. The way forward lies at the intersection of these areas, and it is likely that more than one approach meets all criteria.

The Need for a Process of Ethical Engineering

Many companies would prefer to have an ethical process in place, rather than being subject to regulation or external constraints that require them to hardwire certain outcomes into their products. Relying on a standardized process instead communicates clear expectations to manufacturers, but allows for greater flexibility and thereby enables robust competition by extending to consumers a real choice in the market.

Processes, coupled with responsible record keeping, also provide an opportunity for accountability. Companies may be able to offer compelling legal defenses based on their sincere conformity to processes, thereby demonstrating good faith. If such policies become industry-wide, it provides for a kind of herd legal immunity for manufacturers. Juries might accept as a defense that a manufacturer followed standard procedure, even if liability is at issue.

On the other hand, if manufacturers start losing cases about product liability, or if a jury thinks the manufacturer defendant *should* have programmed their AV to behave another way, it may be hard to preserve the flexibility among manufacturers, and may disincentivize innovation and competition. Manufacturers would prefer to tell a jury, in their defense, that they had made a good faith effort, following standard safety and ethical programming procedures. Standard processes also set a baseline for expectations, so that manufacturers may "go beyond" what is required, further demonstrating good faith.

It is important to note that such a standard process will be orthogonal to any particular moral theory. Different companies or customers may prefer an AV with different priorities or behaviors. Mandating processes rather than outcomes allows for a greater compatibility with a company's values, strategy, and personality.

Such a process could be developed, overseen, and refined by an independent standards body. This body must be respected by the courts, by public opinion, by manufacturers and ethicists. This body could provide a level of guidance that provides a safe legal harbor for manufacturers. Alternately, may be appropriate to constitute a permanent ethics administration inside the Department of Transportation or under NHTSA, analogous to the National Institute of Health's Department of Bioethics, which offers recommendations and analysis to care providers, or *ad hoc* advisory panels composed of engineers and ethicists. It would be difficult politically to vest these bodies with the power to make binding recommendations, but their reports and standards could nevertheless be influential in setting the policy agenda. This kind of guidance can help the industry to move forward, though these processes will be imperfect and continually evolving.

It may be possible soon to develop recommendations for a high-level, general standard for ethical design, for example, modeled on ISO 26262, an industry standard for the safe design of road vehicles. Such a proposal must begin at the general and abstract level. Continued discussions could work to specify and crystallize these recommendations.

Embedding Ethicists with Designers

One concrete recommendation is to embed professional ethicists in the design and engineering process, to ensure they are standardly members of design and engineering project teams. This does not necessarily require ethicists to shadow engineers in their day to day business, but there is a plausible role for ethicists in design decisions. The first valuable contribution of moral philosophers is often to identify when some decision is ethically fraught in a way designers may not appreciate. Manufacturers are already making morally salient decisions without realizing it, and a goal of "maximizing safety" is at least ambiguous and perhaps morally problematic. A careful consideration of the possible tradeoffs can help illuminate an answer that satisfies and is justifiable to the various parties involved, including the broader public. Companies could be incentivized to include professional ethicists in their design process.[19] Ethicists could help to craft a proactive ethics, to "operationalize" the ethics code, and to examine ethical choices that are implicitly made as part of the design process.

This model has been fruitful in the bioethics community for several decades. Many hospitals, for example, convene bioethics boards composed of experts from science, philosophy, and religion, to help them negotiate vexed and morally salient issues that are encountered daily in the hospital and to help them develop policies and processes for moving forward. By encouraging ethicists to work closely with doctors, both disciplines advance in tandem, and both sides benefit from the greater understanding that develops.

Many of the suggestions given here may seem modest. They are an early attempt to satisfy stakeholders with disparate motivations,

goals, worries, disciplinary competencies, and experiences. Clearly, more collaboration is called for, as these discussions are only beginning. Society must match the industry's pace of innovation with our swiftness in examining the attendant ethical problems, with a great sensitivity to the needs of those involved, including the public at large. For now, transparency in the design process and wide-ranging discussions regarding possibilities for ethical programming should be the highest priority.

Works Cited

Arthur, J. (1981). Famine Relief and the Ideal Moral Code. In V. Barry (Ed.), *Applying Ethics*. Belmont, CA: Wadsworth.

Bennett, J. (1998). *The Act Itself*. Oxford University Press.

Bourget, D., & Chalmers, D. (2014). What do philosophers believe? *Philosophical Studies, 170*(3), 465–500.

Consumer Reports. (2014, January). *Black box 101: Understanding event data recorders: All new cars have some form of EDR*. Retrieved from Consumer Reports: http://www.consumerreports.org/cro/2012/10/black-box-101-understanding-event-data-recorders/index.htm

Fehrenbacher, K. (2015, October 16). *How Tesla is ushering in the age of the learning car*. Retrieved from Fortune Magazine: http://fortune.com/2015/10/16/how-tesla-autopilot-learns/

Feinberg, J. (1961). Supererogation and rules. *Ethics, 71*(4), 276–288.

Feinberg, J. (1968). Collective responsibility. *The Journal of Philosophy, 65*(21), 674–688.

Feldman, F. (2002). The good life: A defense of attitudinal hedonism. *Philosophy and Phenomenological Research, 65*, 604–628.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5-15.

Greenberg, A. (2015, July 21). *Hackers Remotely Kill a Jeep on the Highway—With Me in It*. Retrieved from Wired Magazine: https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/

Griffith, E. (2014, August 15). *If driverless cars save lives, where will we get organs?* Retrieved from Fortune Magazine: http://fortune.com/2014/08/15/if-driverless-cars-save-lives-where-will-we-get-organs/

Gunning, D. (n.d.). *Explainable Artificial Intelligence (XAI)*. Retrieved August 30, 2016, from DARPA: http://www.darpa.mil/program/explainable-artificial-intelligence

Hammerschmidt, C. (2015, July 23). *Research Project Tackles Driver Takeover Issue*. Retrieved from EETimes: http://www.eetimes.com/document.asp?doc_id=1327233

Kolitz, D. (2015, December 1). *The lingering effects of NYC's racist city planning*. Retrieved from Hopes and Fears: http://www.hopesandfears.com/hopes/now/politics/216905-the-lingering-effects-of-nyc-racist-city-planning

Lazar, S. (2012). Necessity in Self- Defense and War. *Philosophy & Public Affairs, 40*(1), 3–44.

McDowell, J. (1979). Virtue and reason. *Monist, 62*(3), 331-350.

Miller, D., Sun, A., Johns, M., Ive, H., Sirkin, D., Aich, S., & Ju, W. (2015). Distraction becomes engagement in automated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 59*. SAGE Publications.

National Highway Traffic Safety Administration. (2014). *Fatality Analysis Reporting System (FARS) Encyclopedia*. Retrieved August 28, 2016, from http://www-fars.nhtsa.dot.gov/Main/

National Highway Traffic Safety Administration. (2015). *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey.* National Highway Traffic Safety Administration, US Department of Transportation. Retrieved from National Highway Traffic Safety Administration.

National Highway Traffic Safety Administration. (2015). *The Economic and Societal Impact of Motor Vehicle Crashes, 2010 (Revised).* US Department of Transportation.

National Highway Traffic Safety Administration. (2016). *Early Estimate of Motor Vehicle Traffic Fatalities in 2015.* US Department of Transportation. US Government.

Norcross, A. (1999). Intending and foreseeing death. *Southwest Philosophy Review, 15*(1), 115-123.

Purves, D., Jenkins, R., & Strawser, B. (2015). Autonomous Machines, Moral Judgment, and Acting for the Right Reasons. *Ethical Theory and Moral Practice, 18*(4), 851–872.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy & Public Affairs*, 334-351.

Rachels, J. (1975). Active and passive euthanasia. *The New England journal of medicine, 292*(2), 78–80.

Rawls, J. (1974). Some reasons for the maximin criterion. *The American Economic Review, 64*(2), 141–146.

Rawls, J. (1987). The idea of an overlapping consensus. *Oxford journal of legal studies, 7*(1), 1–25.

Roemer, J. (1998). *Theories of distributive justice.* Harvard University Press.

Safety Research & Strategies, Inc. (2013, November 7). *Toyota Unintended Acceleration and the Big Bowl of "Spaghetti" Code*. Retrieved from The Safety Record Blog: **http://www.safetyresearch.net/blog/articles/toyota-unintended-acceleration-and-big-bowl-"spaghetti"-code**

Shapiro, D. (2016, May 6). *Driver's Ed for Self-Driving Cars: How Our Deep Learning Tech Taught a Car to Drive*. Retrieved from NVIDIA: **https://blogs.nvidia.com/blog/2016/05/06/self-driving-cars-3/**

Solon, O. (2016, July 6). *Lidar: the self-driving technology that could help Tesla avoid another tragedy* . Retrieved from The Guardian: **https://www.theguardian.com/technology/2016/jul/06/lidar-self-driving-technology-tesla-crash-elon-musk**

Spice, B. (2016, May 26). *Carnegie Mellon Transparency Reports Make AI Decision-Making Accountable*. Retrieved from Carnegie Mellon University: CyLab: **https://www.cylab.cmu.edu/news_events/news/2016/carnegie-mellon-transparency-reports-make-ai-decision-making-accountable.html**

Stark, C. (2000). Hypoethical consent and justification. *The Journal of Philosophy, 97*(6), 313–334.

Tesla Motors. (2016, June 30). *A Tragic Loss*. Retrieved from Tesla Motors: **https://www.tesla.com/blog/tragic-loss**

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.

Walch, M., Lange, K., Baumann, M., & Weber, M. (2015). Autonomous Driving: Investigating the Feasibility of Car-Driver Handover Assistance. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 11–18.

Walzer, M. (2008). *Spheres of justice: A defense of pluralism and equality.* Basic Books.

Wernle, B. (2015, May 17). *Mobileye software mimics how the human eye sees*. Retrieved from Automotive News: **http://www.autonews.com/article/20150517/OEM06/305189978/mobileye-software-mimics-how-the-human-eye-sees**

Notes

[1] Estimates published in July 2016 for vehicle deaths in 2015 placed this number at 35,200 deaths (National Highway Traffic Safety Administration, 2016). Historical numbers for 2010 through 2014 range from around 29,000 to 31,000 deaths per year (National Highway Traffic Safety Administration, 2014).

[2] The NHTSA identified driver mistakes as the "critical reason" in 94% of a representative sample of crashes investigated between 2005–2007 (National Highway Traffic Safety Administration, 2015). It is true that the risk of driver failure varies widely between age group, environmental conditions, the driver's mental state, etc. This data may strike some as objectionably coarse-grained. Moreover, some of the most dangerous drivers, such as the very young, the very old, or the inebriated, might be the ones least likely to opt for autonomy.

[3] It is hypothesized that a LIDAR system would have prevented Tesla's deadly May 2016 crash (Solon, 2016).

[4] Moreover, "When quality of life valuations are considered, the total value of societal harm from motor vehicle crashes in 2010 was $836 billion" (National Highway Traffic Safety Administration, 2015).

[5] The term "overlapping consensus" comes from the work of political philosopher John Rawls, who proposed the idea as a way of negotiating political solutions in liberal societies that must accommodate a plurality of conceptions of the good life. By identifying an overlapping consensus among political factions, legislators might find policy positions acceptable to all, even if they disagree about the reasons that justify them (Rawls, 1987).

[6] Appreciation of the vast number of permutations of driving conditions and contexts leads many in industry to doubt that AV will ever be *fully autonomous in all situations*.

[7] The problems here are magnified by the inability of computers to exercise judgment or intuition, which humans can use to guide their behavior in unfamiliar or foreign situations.

[8] This is not to say that technical limitations cannot *inform* policy debates, or that legal regimes can inform and shape technical research moving forward. In this sense, facts about the technical state of the art enter into negotiations about policy. But the kind of technology that exists at a time is not something that can be, e.g., *written into existence* in the way a law can.

[9] Toyota was famously faulted in court by two computer experts who both derided their code as "spaghetti-like," which might have led to their cars' problems with "sudden unintended acceleration" and widespread recalls in 2009–2010 (Safety Research & Strategies, Inc, 2013).

[10] Note, also, that all of the domains of the overlapping consensus should be expected to shift over time. The professional and popular (or "folk") conceptions of morality can both be expected to change (as they more closely approximate the best moral view). And the technical capabilities of manufacturers should also be expected to evolve as the technology becomes more powerful. This is discussed further below.

[11] Both of these examples are from (Rachels, 1975).

[12] Though their explanations differ significantly, a survey of professional philosophers found that 68% of them "accept or lean toward" turning the trolley toward the one person. That number increases to above 70% when the sample is filtered to moral philosophers in particular (Bourget & Chalmers, 2014). Still, the precise comparative value of the deaths that are directly caused versus the deaths that are merely allowed eludes philosophers. If allowing five deaths is worse than killing one person, then what allowing *four* deaths? Three? And so on. It may be hopelessly difficult to fix a value to this comparison that is acceptable to all parties.

[13] The idea of legality-adjustment is inspired by suggestions by other moral philosophers. See, for example, (Feldman, 2002) and (Lazar, 2012).

[14] This analysis is admittedly unrealistically demanding, and will probably be beyond AV for some time. But if future AV will be able to make sense of their surroundings, and make accurate predictions about physics and human behavior, then there is no harm in considering these possibilities now. We may be approaching a future where these capabilities are less implausible, for example, gauging the number of people who stand to be injured, making rough inferences about who is obeying the law, using rough knowledge about the safest crash trajectories.

[15] There are still significant problems with event data recorders (EDRs), the "black boxes" that are used in consumer vehicles to record information about crashes (Consumer Reports, 2014). Storing even thirty seconds of data, when dozens of instruments are sampled hundreds of times per second, requires significant space. Often EDRs can be damaged beyond recovery and so are useless. And juries still struggle to interpret EDR data, for example, trusting driver testimony over the "objective data" in the EDR. NHTSA acknowledges that, "Due to significant limitations however, EDR data should always be used in conjunction with other data sources."

[16] For example, Walch, Lange, Baumann, and Weber, that found that participants in a simulation took about 1.75 seconds to place their hands on the steering wheel of the car after being alerted of a hazard (Walch, Lange, Baumann, & Weber, 2015).

[17] This is one benefit of parking lots: that they ease drivers back into the cognitively complex task of driving on city streets.

[18] Besides defensive countermeasures, manufacturers may also consider "offensive countermeasures" or "hacking back" against an adversary. These are attacks that are intended to affect the operation of the adversary themselves, rather than the victim simply protecting itself. For example, suppose an AV is compromised by a virus, which looks for others AV to spread to. Should those other AV have the freedom, not just to defend themselves, but to disable a compromised and malicious AV? As could be expected, offensive countermeasures are controversial. This is currently a topic of debate and investigation among legal experts. Hacking back could violate anti-hacking laws. There are also worries about false positives leading cars to hack back against innocent actors. Finally, a single AV is likely not properly equipped to instigate its own hacking attack.

[19] See, for example, the Sarbanes-Oxley Act (2002), which required companies to develop ethics standards that guide the conduct of top executives, though presumably these are often anodyne, superficial, or impotent.