Why Ethics Matters for Autonomous Cars

4

Patrick Lin

Content

If motor vehicles are to be truly autonomous and able to operate responsibly on our roads, they will need to replicate – or do better than – the human decision-making process. But some decisions are more than just a mechanical application of traffic laws and plotting a safe path. They seem to require a sense of ethics, and this is a notoriously difficult capability to reduce into algorithms for a computer to follow.

This chapter will explain why ethics matters for autonomous road vehicles, looking at the most urgent area of their programming. Nearly all of this work is still in front of the

P. Lin (✉)
California Polytechnic State University, Philosophy Department, USA
palin@calpoly.edu

M. Maurer et al. (Hrsg.), *Autonomes Fahren*,
DOI 10.1007/978-3-662-45854-9_4,

industry, which is to say that I will mainly raise the questions here and not presume to have any definitive answers at such an early stage of the technology.

A brief note about terminology

I will use "autonomous", "self driving", "driverless", and "robot" interchangeably. These refer primarily to future vehicles that may have the ability to operate without human intervention for extended periods of time and to perform a broad range of actions. I will also use "cars" to refer loosely to all motor vehicles, from a motorcycle to a freight truck; those distinctions do not matter for the discussion here.

4.1 Why ethics matters

To start, let me offer a simple scenario that illustrates the need for ethics in autonomous cars. Imagine in some distant future, your autonomous car encounters this terrible choice: it must either swerve left and strike an eight-year old girl, or swerve right and strike an 80-year old grandmother [33]. Given the car's velocity, either victim would surely be killed on impact. If you do not swerve, both victims will be struck and killed; so there is good reason to think that you ought to swerve one way or another. But what would be the ethically correct decision? If you were programming the self-driving car, how would you instruct it to behave if it ever encountered such a case, as rare as it may be?

Striking the grandmother could be the lesser evil, at least to some eyes. The thinking is that the girl still has her entire life in front of her – a first love, a family of her own, a career, and other adventures and happiness – while the grandmother has already had a full life and her fair share of experiences. Further, the little girl is a moral innocent, more so than just about any adult. We might agree that the grandmother has a right to life and as valuable a life as the little girl's; but nevertheless, there are reasons that seem to weigh in favor of saving the little girl over the grandmother, if an accident is unavoidable. Even the grandmother may insist on her own sacrifice, if she were given the chance to choose.

But either choice is ethically incorrect, at least according to the relevant professional codes of ethics. Among its many pledges, the Institute of Electrical and Electronics Engineers (IEEE), for instance, commits itself and its 430,000+ members "to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression" [23]. Therefore, to treat individuals differently on the basis of their age, when age is not a relevant factor, seems to be exactly the kind of discrimination the IEEE prohibits [18, 33].

Age does not appear to be a relevant factor in our scenario as it might be in, say, casting a young actor to play a child's character in a movie. In that movie scenario, it would be appropriate to reject adult actors for the role. Anyway, a reason to discriminate does not necessarily justify that discrimination, since some reasons may be illegitimate. Even if we point to the disparity of life experiences between the old and the young, that difference isn't automatically an appropriate basis for different treatment.

Discriminating on the basis of age in our crash scenario would seem to be the same evil as discriminating on the basis of race, religion, gender, disability, national origin, and so on, even if we can invent reasons to prefer one such group over another. In Germany – home to many influential automotive companies that are working to develop self-driving technologies – the right to life and human dignity is basic and set forth in the first two articles of the very first chapter in the nation's constitution [9]. So it is difficult to see how German law could even allow a company to create a product that is capable to making such a horrific and apparently illegal choice. The United States similarly strives to offer equal protection to all persons, such as stipulated in the fourteenth amendment of its constitution.

If we cannot ethically choose a path forward, then what ought to be done? One solution is to refuse to make a swerve decision, allowing both victims to be struck; but this seems much worse than having only one victim die, even if we are prejudiced against her. Anyway, we can force a decision by modifying the scenario: assume that 10 or 100 other pedestrians would die, if the car continued forward; and swerving would again result in only a single death.

Another solution could be to arbitrarily and unpredictably choose a path, without prejudice to either person [34]. But this too seems ethically troubling, in that we are choosing between lives without any deliberation at all – to leave it to chance, when there are potentially some reasons to prefer one over the other, as distasteful and uncomfortable as those reasons may be. This is a dilemma that is not easily solvable and therefore points to a need for ethics in developing autonomous cars.

4.1.1 Beyond crash-avoidance

Many readers may object right away that the dilemma above (and others that follow) will never occur with autonomous cars. It may be suggested that future cars need not confront hard ethical choices, that simply stopping the car or handing control back to the human operator is the easy path around ethics. But I will contend here that braking and relinquishing control will not always be enough. Those solutions may be the best we have today, but if automated cars are to ever operate more broadly outside of limited highway environments, they will need more response-options.

Current research already makes this case as a matter of physics [12, 13], but we can also make a case from commonsense. Many ordinary scenarios exist today in which braking is not the best or safest move, whether by human or self-driving car. A wet road or a tailgater, for instance, may make it dangerous to slam the brakes, as opposed to some other action such as steering around the obstacle or simply through it, if it is a small object. Today, the most advanced self-driving cars cannot detect small objects such as squirrels [7]; therefore, they presumably cannot also detect squirrel-sized rocks, potholes, kittens, and other small but consequential hazards can cause equipment failure, such as tire blowouts or sensor errors, or deviations from a safe path.

In these and many other cases, there may not be enough time to hand control back to the driver. Some simulation experiments suggest that human drivers need up to 40 seconds to regain situation awareness, depending on the distracting activity, e. g., reading or napping – far longer than the 1–2 seconds of reaction time required for typical accident scenarios [38, 18]. This means that the car must be responsible for making decisions when it is unreasonable to expect a timely transfer of control back to the human, and again braking might not be the most responsible action.

One possible reply is that, while imperfect, braking could successfully avoid the majority of emergency situations a robot car may find itself it, even if it regrettably makes things worse in a small number of cases. The benefits far outweigh the risks, presumably, and the numbers speak for themselves. Or do they? I will discuss the dangers of morality by math throughout this chapter.

Braking and other responses in the service of crash-avoidance won't be enough, because crash-avoidance is not enough. Some accidents are unavoidable – such as when an animal or pedestrian darts out in front of your moving car – and therefore autonomous cars will need to engage in crash-*optimization* as well. Optimizing crashes means to choose the course of action that will likely lead to the least amount of harm, and this could mean a forced choice between two evils, for instance, choosing to strike either the eight-year old girl or the 80-year old grandmother in my first scenario above.

4.1.2 Crash-optimization means targeting

There may be reasons, by the way, to prefer choosing to run over the eight-year old girl that I have not yet mentioned. If the autonomous car were most interested in protecting its own occupants, then it would make sense to choose a collision with the lightest object possible (the girl). If the choice were between two vehicles, then the car should be programmed to prefer striking a lighter vehicle (such as a Mini Cooper or motorcycle) than a heavier one (such as a sports utility vehicle (SUV) or truck) in an adjacent lane [18, 34].

On the other hand, if the car were charged with protecting other drivers and pedestrians over its own occupants – not an unreasonable imperative – then it should be programmed to prefer a collision with the heavier vehicle than the lighter one. If vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications are rolled out (or V2X to refer to both), or if an autonomous car can identify the specific models of other cars on the road, then it seems to make sense to collide with a safer vehicle (such as a Volvo SUV that has a reputation for safety) over a car not known for crash-safety (such as a Ford Pinto that's prone to exploding upon impact).

This strategy may be both legally and ethically better than the previous one of jealously protecting the car's own occupants. It could minimize lawsuits, because any injury to others would be less severe. Also, because the driver is the one who introduced the risk to society – operating an autonomous vehicle on public roads – the driver may be legally obligated,

or at least morally obligated, to absorb the brunt of any harm, at least when squared off against pedestrians, bicycles, and perhaps lighter vehicles.

The ethical point here, however, is that no matter which strategy is adopted by an original equipment manufacturer (OEM), i. e., auto manufacturer, programming a car to choose a collision with any particular kind of object over another very much resembles a *targeting* algorithm [33]. Somewhat related to the military sense of selecting targets, crash-optimization algorithms may involve the deliberate and systematic discrimination of, say, large vehicles or Volvos to collide into. The owners or operators of these targeted vehicles bear this burden through no fault of their own, other than perhaps that they care about safety or need an SUV to transport a large family.

4.1.3 Beyond harm

The problem is starkly highlighted by the following scenario [15, 16, 17, 34]: Again, imagine that an autonomous car is facing an imminent crash, but it could select one of two targets in adjacent lanes to swerve into: either a motorcyclist who is wearing a helmet, or a motorcyclist who is not. It probably doesn't matter much to the safety of the car itself or its occupants whether the motorcyclist is wearing a helmet; the impact of a helmet into a car window doesn't introduce that much more risk that the autonomous car should want to avoid it over anything else. But it matters a lot to the motorcyclist whether s/he is wearing a helmet: the one without a helmet would probably not survive such a collision. Therefore, in this dreadful scenario, it seems reasonable to program a good autonomous car to swerve into the motorcyclist with the helmet.

But how well is justice and public policy served by this crash-optimization design? Motorcyclists who wear helmets are essentially being penalized and discriminated against for their responsible decision to wear a helmet. This may encourage some motorcyclists to not wear helmets, in order to avoid targeting by autonomous cars. Likewise, in the previous scenario, sales may decline for automotive brands known for safety, such as Volvo and Mercedes Benz, insofar as customers want to avoid being the preferred targets of crash-optimization systems.

Some readers may want to argue that the motorcyclist without a helmet ought to be targeted, for instance, because he has acted recklessly and therefore is more deserving of harm. Even if that's the correct design, notice that we are again moving beyond harm in making crash-optimization decisions. We're still talking about justice and other such ethical considerations, and that's the point: it's not just a numbers game.

Programmers in such scenarios, as rare as they may be, would need to design cost-functions – algorithms that assign and calculate the expected costs of various possible options, selecting the one with the lowest costs – that potentially determine who gets to live and who gets to die. And this is fundamentally an ethics problem, one that demands much more care and transparency in reasoning than seems currently offered. Indeed, it is difficult to imagine a weightier and more profoundly serious decision

a programmer would ever have to make. Yet, there is little discussion about this core issue to date.

4.2 Scenarios that implicate ethics

In addition to the ones posited above, there are many actual and hypothetical scenarios that involve judgments about ethics. I will describe some here to show how ordinary assumptions in ethics can be challenged.

4.2.1 The deer

Though difficult to quantify due to inconsistent and under-reporting, experts estimate that more than a million car accidents per year in the US are caused by deer [6, 48]. Many, if not most, drivers have been startled by an unexpected animal on the road, a dangerous situation for both parties. Deconstructing a typical accident, or near-accident, involving an animal illustrates the complexity of the decisions facing the driver [30]. While all this happens within seconds – not enough time for careful deliberations by human drivers – an autonomous car could have the virtue of a (presumably) thoughtful decision-making script to very quickly react in an optimal way. If it is able to account for the many variables, then it ought to, for the most informed decision possible.

First, suppose an object appears on the road directly in front of a car in autonomous mode. Is there time to reasonably hand control back to the human behind the wheel? (Probably not.) If not, is there time to stop the car? Would the car need to brake hard, or would moderate braking be sufficient? The decision to brake depends, again, on road conditions and whether a tailgater (such as a big-rig truck) is behind you, including its speed to determine the severity of a possible rear-end collision.

Second, what is the object? Is it an animal, a person, or something else? If it is an animal, are some animals permissible to run over? It may be safer to continue ahead and strike a squirrel, for instance, than to violently swerve around it and risk losing control of the car. However, larger animals, such as deer and cows, are more likely to cause serious damage to the car and injuries to occupants than a spun-out car. Other animals, still, have special places in our hearts and should be avoided if possible, such as pet dogs and cats.

Third, if the car should get out of the way – either in conjunction with braking or not – should it swerve to the left or to the right? In the US and other nations in which drivers must stay on the right side of the road, turning to the right may mean driving off the road, potentially into a ditch or a tree. Not only could harm to the car and occupants be likely, but it also matters how many occupants are in the car. The decision to drive into an embankment seems different when only one adult driver is in the car, than when several children are inside too.

On the other hand, turning to the left may mean driving into an opposite lane, potentially into a head-on collision with incoming vehicles. If such a collision is unavoidable, then it

matters what kind of vehicle we would crash into (e. g., is it a compact car or SUV?), how heavy incoming traffic is (e. g., would more than one vehicle be involved?), how many persons may be involved (e. g., are there children in the other car?). Of course, here we are assuming perfect sensing and V2X communications that can help answer these questions. If we cannot answer the questions, then we face a possibly large unknown risk, which makes driving into incoming traffic perhaps the worst option available.

Other factors relevant to the decision-points above include: the road-shoulder type (paved, gravel, none, etc.), the condition of the car's tires and brakes, whether the car's occupants are seat-belted, whether the car is transporting dangerous cargo that could spill or explode, proximity to hospital or emergency rescue, damage to property such as houses and buildings, and more. These variables influence the probability of an accident as well as expected harm, both of which are needed in selecting the best course of action.

From this short analysis of a typical crash (or possible crash) with an animal, we can already see a daunting number of factors to account for. Sensing technologies today cannot answer some or many of the questions above, but it is already unclear that braking should be the safest default option – as a proxy for the most ethical option – given these uncertain conditions, all things considered. Automated cars today can already detect whether there is oncoming traffic in the opposite lane. Therefore, it is at least possible that they can be programmed to maneuver slightly into the incoming lane under some conditions, e. g., when there are no incoming cars and when it may be dangerous to slam on the brakes.

Whether or not sensing technologies will improve enough to deliver answers to our questions above, a programmer or OEM would still need to assign costs or weights to various actions and objects as best as they can. Yet these values are not intrinsic to or discoverable by science or engineering. Values are something that we humans must stipulate and ideally agree upon. In constructing algorithms to control an autonomous car, ethics is already implied in the design process. Any decision that involves a tradeoff such as to strike object x instead of object y requires a value-judgment about the wisdom of the tradeoff, that is, the relative weights of x and y. And the design process can be made better by recognizing the ethical implications and by engaging the broader community to ensure that those values are represented correctly or at least transparently. Working in a moral bubble is less likely to deliver results that are acceptable to society.

Again, in a real-world accident today, a human driver usually has neither the time nor the information needed to make the most ethical or least harmful decisions. A person who is startled by a small animal on an otherwise uneventful drive may very well react poorly. He might drive into oncoming traffic and kill a family, or oversteer into a ditch and to his own death. Neither of these results, however, is likely to lead to criminal prosecution by themselves, since there was no forethought, malice, negligence, or bad intent in making a forced, split-second reaction. But the programmer and OEM do not operate under the sanctuary of reasonable instincts; they make potentially life-and-death decisions under no truly urgent time-constraint and therefore incur the responsibility of making better decisions than human drivers reacting reflexively in surprise situations.

4.2.2 Self-sacrifice

As we can see, real-world accidents can be very complicated. In philosophy and ethics, a familiar method is to simplify the issues through hypothetical scenarios, otherwise known as "thought-experiments." This is similar to everyday science experiments in which researchers create unusual conditions to isolate and test desired variables, such as sending spiders into outer space to see how micro-gravity affects their ability to spin webs. It is not a good objection to those experiments to say that no spiders exist naturally in space; that misses the point of the experiment.

Likewise, it is no objection to our hypothetical examples that they are outlandish and unlikely to happen in the real world, such as a car that can distinguish an eight-year old from an 80-year old (though with improving biometrics, facial recognition technologies, and linked databases, this doesn't seem impossible). Our thought-experiments are still useful in drawing out certain ethical intuitions and principles we want to test.

With that understanding, we can devise hypothetical scenarios to see that reasonable ethical principles can lead to controversial results in the context of autonomous driving. Digging into a standard philosophical toolbox for help with ethical dilemmas, one of the first principles we might reach for is consequentialism: that the right thing to do is whatever leads to the best results, especially in quantified terms [44]. As it applies here, consequentialism suggests that we should strive to minimize harm and maximize whatever it is that matters, such as, the number of happy lives.

In this thought-experiment, your future autonomous car is driving you on a narrow road, alongside a cliff. No one and no technology could foresee that a school bus with 28 children would appear around the corner, partially in your lane [29, 36]. Your car calculates that crash is imminent; given the velocities and distance, there is no possible action that can avoid harming you. What should your robot car do?

A good, standard-issue consequentialist would want to optimize results, that is, maximize the number of happy lives and minimize harm. Assuming that all lives in this scenario are more or less equally happy – for instance, there's no super-happy or super-depressed person, and no very important person who has unusual influence over the welfare of others – they would each count for about the same in our moral calculation. As you like, we may either ignore or account for the issue of whether there is extra value in the life of innocent child who has more years of happiness ahead of her than an average adult; that doesn't matter much for this scenario.

The robot car's two main choices seem to be: (1) to slam on the brakes and crash into the bus, risking everyone's lives, or (2) to drive off the cliff, sparing the lives of everyone on the bus. Performing a quick expected-utility calculation, if the odds of death to each person (including the adult bus driver) in the accident averaged more than one in 30, then colliding into the bus would yield the expected result of more than one death, up to all 30 persons. (Let's say the actual odds are one in three, which gives an expected result of 10 deaths.) If driving off a cliff meant certain death, or the odds of one in one, then the expected result of that would be exactly one death (your own) and no more. The right consequen-

tialist decision for the robot car – if all we care about is maximizing lives and minimizing deaths – is apparently to drive off the cliff and sacrifice the driver, since it is better that only one person should die rather than more than one, especially 10 or all 30 persons.

This decision would likely be different if, instead of a school bus, your robot car were about to collide with another passenger car carrying only one person. Given the same average odds of death, one in three, the expected number of deaths in a collision would only be 0.67, while the expected number of deaths in driving off a cliff remains at one. In that case, the right consequentialist decision would be to allow the accident to occur, as long as the average odds of death are less than one in two. If, instead of another vehicle, your car were about to collide with a deer, then the decision to stay on the road, despite an ensuing accident, would be even more obvious insofar as we value a deer's life less than a human life.

Back to the school-bus scenario, programming an autonomous car with a consequentialist framework for ethics would seem to imply your sacrifice. But what is most striking about this case might not even be your death or the moral mathematics: if you were in a manually driven car today, driving off the cliff might still be the most ethical choice you could make, so perhaps you would choose certain death anyway, had you the time to consider the options. However, it is one thing for you to willingly make that decision of sacrifice yourself, and quite another matter for a machine to make that decision without your consent or foreknowledge that self-sacrifice was even a possibility. That is, there is an astonishing lack of transparency and therefore consent in such a grave decision, one of the most important that can be made about one's life – perhaps noble if voluntary, but criminal if not.

Thus, reasonable ethical principles – e.g., aiming to save the greatest number of lives – can be stressed in the context of autonomous driving. An operator of an autonomous vehicle, rightly or not, may very well value his own life over that of everyone else's, even that of 29 others; or he may even explicitly reject consequentialism. Even if consequentialism is the best ethical theory and the car's moral calculations are correct, the problem may not be with the ethics but with a lack of discussion about ethics. Industry, therefore, may do well to have such a discussion and set expectations with the public. Users – and news headlines – may likely be more forgiving if it is explained in advance that self-sacrifice may be a justified feature, not a bug.

4.2.3 Ducking harm

Other ethical principles can create dilemmas, too. It is generally uncontroversial that, if you can easily avoid harm to yourself, then you should do it. Indeed, it may be morally required that you save yourself when possible, if your life is intrinsically valuable or worth protecting; and it is at least extrinsically valuable if you had a dependent family. Auto manufacturers or OEMs seem to take this principle for granted as well: if an autonomous car can easily avoid a crash, e. g., by braking or swerving, then it should. No ethical problem here – or is there?

In another thought-experiment [15, 18, 33], your robotic car is stopped at an intersection and waits patiently for the children who are crossing in front of you. Your car detects a pickup truck coming up behind you, about to cause a rear-end collision with you. The crash would likely damage your car to some degree and perhaps cause minor injury to you, such as whiplash, but certainly not death. To avoid this harm, your car is programmed to dash out of the way, if it can do so safely. In this case, your car can easily turn right at the intersection and avoid the rear-end collision. It follows this programming, but in doing so, it clears a path for the truck to continue through the intersection, killing a couple children and seriously injuring others.

Was this the correct way to program an autonomous car? In most cases of an impending rear-end collision, probably yes. But in this particular case, the design decision meant saving you from minor injury at the expense of serious injury and death of several children, and this hardly seems to be the right choice. In an important respect, you (or the car) are responsible for their deaths: you (or the car) killed the children by removing an obstruction that prevented harm from falling upon them, just as you would be responsible for a person's death if you removed a shield he was holding in front of a stream of gunfire. And killing innocent people has legal and moral ramifications.

As with the self-sacrifice scenario above, it might be that in the same situation today, in a human-driven car, you would make the same decision to save yourself from injury, if you were to see a fast-approaching vehicle about to slam into you. That is, the result might not change if a human made the on-the-spot decision. But, again, it is one thing to make such a judgment in the panic of the moment, but another less forgivable thing for a programmer – far removed from the scene and a year or more in advance – to create a cost-function that resulted in these deaths. Either the programmer did so deliberately, or she did it unintentionally, unaware that this was a possibility. If the former, then this could be construed as premeditated homicide; and if the latter, gross negligence.

Either way is very bad for the programmer and perhaps an inherent risk in the business, when one attempts to replicate human decision-making in a broad range of dynamic scenarios. Sometimes, an autonomous car may be faced with a "no-win" scenario, putting the programmer in a difficult but all too real position. To mitigate this risk, industry may do well to set expectations not only with users but also with broader society, educating them that they could also become victims even if not operating or in a robot car, and that perhaps this is justified by a greater public or overall good.

4.2.4 Trolley problems

One of the most iconic thought-experiments in ethics is the trolley problem [4, 8, 11, 47], and this is one that may now occur in the real world, if autonomous vehicles come to be. Indeed, driverless trains are already operating in dozens of cities worldwide and could bring this scene to life [24]. The classical dilemma involves a runaway trolley (or train) that is about to run over and kill five unaware people standing on the tracks. Looking at the scene

from the outside, you find yourself standing next to a switch: if you pull the switch, you can shunt the train to a right-hand set of tracks, thereby saving the five individuals on the track. Unfortunately, there is one person standing on the right-hand set of tracks who would then be killed. What is the right decision?

The "correct" decision continues to be a subject of much debate in philosophy. Both answers seem reasonable and defensible. A consequentialist might justify switching the tracks to save five people, even at the regrettable expense of one. But a non-consequentialist, someone who considers more than just the math or results, might object on the grounds that switching tracks constitutes an act of killing (the one person), while doing nothing is merely allowing someone to die (the five individuals); and that it is morally and legally worse to kill than to let die.

Killing implies that you are directly responsible for a person's death: had you not done what you did, the person would have lived. Letting die, however, involves much less responsibility on your part, if any, since some causal process was already underway that was not initiated or otherwise controlled by you. The question of whether it is worse to kill than to let die is also subject to debate in philosophy. But let us bracket that for the moment, as a final answer is not necessary for our discussion, only that it is reasonable to believe that proposition.

Adapting the trolley problem to the technology at hand, let us suppose that you are driving an autonomous car in manual mode; you are in control. Either intentionally or not – you could be homicidal or simply inattentive – you are about to run over and kill five pedestrians. Your car's crash-avoidance system detects the possible accident and activates, forcibly taking control of the car from your hands. To avoid this disaster, it swerves in the only direction it can, let's say to the right. But on the right is a single pedestrian who is unfortunately killed.

Was this the right decision for your car to make? Again, a consequentialist would say yes: it is better that only one person dies than five. But a non-consequentialist might appeal to a moral distinction between killing and letting die, and this matters to OEMs for liability reasons. If the car does not wrestle control from the human driver, then it (and the OEM) would perhaps not be responsible for the deaths of the five pedestrians while you were driving the car; it is merely letting those victims die. But if the car does take control and make a decision that results in the death of a person, then it (and the OEM) becomes responsible for killing a person.

As with the trolley problem, either choice seems defensible. Results do matter, so it is not ridiculous to think that the car should be programmed to act and save lives, even at the expense of a fewer number of lives. Yet it also seems reasonable to think that killing is worse than letting die, especially in the eyes of the law. What I want to highlight here is not so much the answer but the process of deliberation that points us toward one answer over another. To the extent that there could be many acceptable answers to any given ethical dilemma, how well one answer can be defended is crucial toward supporting that answer over others.

Industry again would do well to set expectations by debating and explaining in advance its reasoning behind key algorithms that could result in life or death. Transparency, or showing one's math, is an important part of doing ethics, not just the answer itself.

4.3 Next steps

Notice that the ethical issues discussed in this paper do not depend on technology errors, poor maintenance, improper servicing, security vulnerabilities, or other failings – and all those will occur too. No complex technology we have created has been infallible. Even industries with money directly at stake have not solved this problem. For instance, bank ATMs continue to make headlines when they hemorrhage cash – tens of thousands of dollars more than the account holder actually has – because of software glitches alone [2, 10], never mind hacking. And just about every computing device we have created has been hacked or is hackable, including neural implants and military systems [3, 28].

These vulnerabilities and errors certainly can cause harm in the context of autonomous cars, and it would be unethically irresponsible to not safeguard against them where we can. Putting these technology issues aside and even assuming that perfect technology is available, there are still many other safety and ethical questions to worry about, such as the programming issues above.

4.3.1 Broader ethical issues

But programming is only one of many areas to reflect upon as society begins to adopt autonomous driving technologies. Assigning legal and moral responsibility for crashes is a popular topic already [1, 14, 20, 22, 49, 51]. Here are a few others, as part of a much longer list of possible questions:

Does it matter to ethics if a car is publicly owned, for instance, a city bus or fire truck? The owner of a robot car may reasonably expect that its property "owes allegiance" to the owner and should value his or her life more than anonymous pedestrians and drivers. But a publicly owned automated vehicle might not have that obligation, and this can change moral calculations. Even for privately owned autonomous vehicles, the occupants arguably should bear more or all of the risk, since they are the ones introducing the machine into public spaces in the first place.

Do robot cars present an existential threat to the insurance industry? Some believe that ultra-safe cars that can avoid most or all accidents will mean that many insurance companies will go bankrupt, since there would be no or very little risk to insure against [40, 52]. But things could go the other way too: We could see *mega*-accidents as cars are networked together and vulnerable to wireless hacking – something like the stock market's "flash crash" in 2010 [5]. What can the insurance industry do to protect itself while not getting in the way of the technology, which holds immense benefits?

How susceptible would robot cars be to hacking? So far, just about every computing device we have created has been hacked. If authorities and owners (e. g., rental car company) are able to remotely take control of a car – which is reportedly under development for law enforcement in the European Union [50] – this offers an easy path for cyber-carjackers. If under attack, whether a hijacking or ordinary break-in, what should the car do: speed

away, alert the police, remain at the crime scene to preserve evidence, or maybe defend itself?

For a future suite of in-car apps, as well as sensors and persistent GPS/tracking, can we safeguard personal information, or do we resign ourselves to a world with disappearing privacy rights [27]? To the extent that online services bring online advertising, we could see new, insidious advertising schemes that may allow third-party advertisers to have some influence on the autonomous car's route selection, e.g., steering the car past their businesses [32].

What kinds of abuse might we see with autonomous cars? If the cars drive too conservatively, they may become a traffic hazard or trigger road-rage in human drivers with less patience [26, 42]. If the crash-avoidance system of a robot car is generally known, then other drivers may be tempted to "game" it, e. g., by cutting in front of it, knowing that the automated car will slow down or swerve to avoid an accident. If those cars can safely drive us home in a fully-auto mode, that may encourage a culture of more alcohol consumption, since we won't need to worry so much about drunk-driving.

More distant concerns include: How will law-abiding robot cars affect city revenue, which often depends on traffic fines imposed against law-breaking human drivers? Inasmuch as many organ transplants come from car-accident victims, how will society manage a declining and already insufficient supply of donated organs [41]?

Older-model autonomous cars may be unable to communicate with later models or future road infrastructure. How do we get those legacy models – which may be less safe, in addition to incompatible with newer technology – off the roads [45]? Since 2009, Microsoft has been trying to kill off its Windows XP operating system [39], a much less expensive investment than an autonomous car; but many users still refuse to relinquish it, including for critical military systems [37, 46]. This is a great security risk since Microsoft will no longer offer software patches for the operating system.

4.3.2 Conclusions

We don't really know what our robot-car future will look like, but we can already see that much work needs to be done. Part of the problem is our lack of imagination. Technology policy expert Peter W. Singer observed, "We are still at the 'horseless carriage' stage of this technology, describing these technologies as what they are not, rather than wrestling with what they truly are" [43].

As it applies here, robots aren't merely replacing human drivers, just as human drivers in the first automobiles weren't simply replacing horses: that would like mistaking electricity as merely a replacement for candles. The impact of automating transportation will change society in radical ways, and technology seems to be accelerating. As Singer puts it, "Yes, Moore's Law is operative, but so is Murphy's Law" [43]. When technology goes wrong – and it will – thinking in advance about ethical design and policies can help guide us responsibility into the unknown.

In future autonomous cars, crash-avoidance features alone won't be enough. An accident may be unavoidable as a matter of physics [12, 13], especially as autonomous cars make their way onto city streets [19, 21, 25], a more dynamic environment than highways. It also could be too dangerous to slam on the brakes, or not enough time to hand control back to the unaware human driver, assuming there's a human in the vehicle at all. Technology errors, misaligned sensors, malicious actors, bad weather, and bad luck can also contribute to imminent collisions. Therefore, robot cars will also need to have crash-optimization strategies that are thoughtful about ethics.

If ethics is ignored and the robotic car behaves badly, a powerful case could be made that auto manufacturers were negligent in the design of their product, and that opens them up to tremendous legal liability, should such an event happen. Today, we see activists campaigning against "killer" military robots that don't yet exist, partly on the grounds that machines should never be empowered to make life-and-death decisions [31, 35]. It's not outside the realm of possibility to think that the same precautionary backlash won't happen to the autonomous car industry, if industry doesn't appear to be taking ethics seriously.

The larger challenge, though, isn't just about thinking through ethical dilemmas. It's also about setting accurate expectations with users and the general public who might find themselves surprised in bad ways by autonomous cars; and expectations matter for market acceptance and adoption. Whatever answer to an ethical dilemma that industry might lean towards will not be satisfying to everyone. Ethics and expectations are challenges common to all automotive manufacturers and tier-one suppliers who want to play in this emerging field, not just particular companies.

Automated cars promise great benefits and unintended effects that are difficult to predict, and the technology is coming either way. Change is inescapable and not necessarily a bad thing in itself. But major disruptions and new harms should be anticipated and avoided where possible. That is the role of ethics in innovation policy: it can pave the way for a better future while enabling beneficial technologies. Without looking at ethics, we are driving with one eye closed.

References

1. Anderson, J., Kalra, N., Stanley, K., Sorensen, P., Samaras, C., Oluwatola, O.: Autonomous vehicle technology: a guide for policymakers. Report by RAND Corporation. http://www.rand.org/pubs/research_reports/RR443-1.html (2014). Accessed 8 July 2014
2. Associated Press. ATM 'glitch' gives $37,000 to lucky homeless man. Daily Mail. http://www.dailymail.co.uk/news/article-2596977/ATM-glitch-gives-OVER-37-000-homeless-man-cash-requested-140.html (2014). Accessed 8 July 2014
3. Baldor, L.: China hacked the Pentagon to get weapons data. Talking Points Memo. http://talkingpointsmemo.com/news/china-hacked-the-pentagon-to-get-weapons-programs-data (2013). Accessed 8 July 2014
4. Cathcart, T.: The Trolley Problem, or Would You Throw the Fat Guy Off the Bridge? Workman Publishing Company, New York (2013)

5. Commodity Futures Trading Commission and the Securities and Exchange Commission: Findings regarding the market events of May 6, 2010. CFTC and SEC, Washington DC. http://www.sec.gov/news/studies/2010/marketevents-report.pdf (2010). Accessed 8 July 2014
6. Curtis, P. and Hedlund, J.: Reducing deer-vehicle crashes. Report funded by the Insurance Institute for Highway Safety. Cornell University, Ithaca. http://wildlifecontrol.info/pubs/Documents/Deer/Deer-Vehicle_factsheet1.pdf (2005). Accessed 8 July 2014
7. Davies, A.: Avoiding squirrels and other things Google's robot car can't do. Wired. http://www.wired.com/2014/05/google-self-driving-car-can-cant/ (2014). Accessed 8 July 2014
8. Edmonds, D.: Would You Kill the Fat Man? The Trolley Problem and What Your Answer Tells Us About Right and Wrong. Princeton University Press, Princeton (2014)
9. Federal Ministry of Justice and Consumer Protection: Basic Law for the Federal Republic of Germany. http://www.gesetze-im-internet.de/englisch_gg/englisch_gg.html (2014). Accessed 8 July 2014
10. Floro, Z.: Man goes on a casino bender after an ATM let him take out unlimited cash. Business Insider. http://www.businessinsider.com/atm-gives-unlimited-cash-2012-6 (2012). Accessed 8 July 2014
11. Foot, P.: The problem of abortion and the doctrine of the double effect. Oxford Review 5, 5–15 (1967)
12. Fraichard, T.: Will the driver seat ever be empty? Report funded by Inria. http://hal.inria.fr/hal-00965176 (2014). Accessed 8 July 2014
13. Fraichard, T., and Asama, H.: Inevitable collision states: a step towards safer robots? Advanced Robotics 18(10), 1001–1024 (2004)
14. Garza, A.: 'Look Ma, no hands!': wrinkles and wrecks in the age of autonomous vehicles. New England Law Review 46(3), 581–616 (2012)
15. Goodall, N.J.: Autonomous car ethics. Interview with CBC radio. http://www.cbc.ca/spark/blog/2014/04/13/autonomous-car-ethics/ (2014). Accessed 8 July 2014
16. Goodall, N. J.: Machine ethics and automated vehicles. In: Meyer, G. and Beiker, S. (eds.) Road Vehicle Automation. Springer, Cham (2014)
17. Goodall, N. J.: Ethical decision making during automated vehicle crashes. Transportation Research Record: Journal of the Transportation Research Board (forthcoming)
18. Goodall, N. J.: Vehicle automation and the duty to act. In: Proceedings of the 21st World Congress on Intelligent Transport Systems, 7–11 September 2014, Detroit, Michigan (forthcoming)
19. Google: The latest chapter for the self-driving car: mastering city street driving. http://googleblog.blogspot.co.at/2014/04/the-latest-chapter-for-self-driving-car.html (2014). Accessed 8 July 2014
20. Gurney, J.: Sue my car, not me: products liability and accidents involving autonomous vehicles. Journal of Law, Technology and Policy 2, 247–277 (2013)
21. Hern, A.: Self-driving cars face a long and winding road to success. The Guardian. http://www.theguardian.com/technology/2014/may/28/self-driving-cars-google-success (2014). Accessed 8 July 2014
22. Hevelke, A. and Nida-Rumelin, J.: Responsibility for crashes of autonomous vehicles: an ethical analysis. Science and Engineering Ethics (2014). doi: 10.1007/s11948-014-9565-5
23. IEEE: IEEE code of ethics. http://www.ieee.org/about/corporate/governance/p7-8.html (2014). Accessed 8 July 2014
24. International Association of Public Transport: Observatory of automated metros. http://metroautomation.org/ (2014). Accessed 8 July 2014
25. Jaffe, E.: The first look at how Google's self-driving car handles city streets. The Atlantic/CityLab. http://www.citylab.com/tech/2014/04/first-look-how-googles-self-driving-car-handles-city-streets/8977/ (2014). Accessed 8 July 2014

26. KPMG: Self-driving cars: are we ready? Report by KPMG. KPMG, Chicago. http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/self-driving-cars-are-we-ready.pdf (2013). Accessed 8 July 2014
27. Lee, T.: Self-driving cars are a privacy nightmare. And it's totally worth it. The Washington Post. http://www.washingtonpost.com/blogs/wonkblog/wp/2013/05/21/self-driving-cars-are-a-privacy-nightmare-and-its-totally-worth-it/ (2013). Accessed 8 July 2014
28. Leggett, H.: The new hacking frontier: your brain? Wired. http://www.wired.com/2009/07/neurosecurity/ (2009). Accessed 8 July 2014
29. Lin, P.: The ethics of saving lives with autonomous cars is far murkier than you think. Wired. http://www.wired.com/2013/07/the-surprising-ethics-of-robot-cars/ (2013). Accessed 8 July 2014
30. Lin, P.: The ethics of autonomous cars. The Atlantic. http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/ (2013). Accessed 8 July 2014
31. Lin, P.: Why the drone wars matter for automated cars. Lecture presented at Proceedings in Automated Driving, Stanford Law School, 12 December 2013. http://stanford.io/1jeIQuw. Accessed 8 July 2014
32. Lin, P.: What if your autonomous car keeps routing you past Krispy Kreme? The Atlantic. http://www.theatlantic.com/technology/archive/2014/01/what-if-your-autonomous-car-keeps-routing-you-past-krispy-kreme/283221/ (2014). Accessed 8 July 2014
33. Lin, P.: Ethics and autonomous cars: why ethics matters, and how to think about it. Lecture presented at Daimler and Benz Foundation's Villa Ladenburg Project, Monterey, California, 21 February 2014
34. Lin, P.: The robot car of tomorrow might just be programmed to hit you. Wired. http://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/ (2014). Accessed 8 July 2014
35. Lin, P., Bekey, G., Abney, K.: Autonomous military robotics: risk, ethics, and design. Report funded by the US Office of Naval Research. California Polytechnic State University, San Luis Obispo. http://ethics.calpoly.edu/ONR_report.pdf (2008). Accessed 8 July 2014
36. Marcus, G.: Moral machines. New Yorker. http://www.newyorker.com/online/blogs/newsdesk/2012/11/google-driverless-car-morality.html (2012). Accessed 8 July 2014
37. McCabe, R.: Navy, others still struggling to ditch Windows XP. The Virginian-Pilot. http://hamptonroads.com/2014/06/navy-others-still-struggle-ditching-windows-xp (2014). Accessed 8 July 2014
38. Merat, N., Jamson, H., Lai F., and Carsten, O.: Human factors of highly automated driving: results from the EASY and CityMobil projects. In: Meyer, G. and Beiker, S. (eds) Road Vehicle Automation. Springer, Cham (2014)
39. Microsoft: Windows lifecycle fact sheet. http://windows.microsoft.com/en-us/windows/lifecycle (2014). Accessed 8 July 2014
40. Mui, C. and Carroll, P.: Driverless Cars: Trillions Are Up For Grabs. Cornerloft Press, Seattle (2013)
41. Park, A.: Driverless cars to kill organ transplantation. DriverlessCarHQ. https://web.archive.org/web/20120626151201/http://www.driverlesscarhq.com/transplants/ (2012). Accessed 8 July 2014
42. Roach, J.: Road rage at driverless cars? It's possible. NBC News. http://futureoftech-discuss.nbcnews.com/_news/2012/01/20/10201865-road-rage-at-driverless-cars-its-possible (2012). Accessed 8 July 2014
43. Singer, P.W.: The robotic revolution. Brookings Institution. http://www.brookings.edu/research/opinions/2012/12/11-robotics-military-singer (2012). Accessed 8 July 2014

44. Sinnott-Armstrong, W.: Consequentialism. Stanford Encyclopedia of Philosophy. http://plato.stanford.edu/entries/consequentialism/ (2011). Accessed 8 July 2014
45. Smith, B.W.: Planning for the obsolescence of technologies not yet invented. Stanford Law School's Center for Internet and Society. http://cyberlaw.stanford.edu/blog/2013/10/planning-obsolescence-technologies-not-yet-invented (2013). Accessed 8 July 2014
46. Sternstein, A.: Why Feds are still buying IT that works with Windows XP. Nextgov. http://www.nextgov.com/cio-briefing/2014/04/why-feds-are-still-buying-it-works-windows-xp/81667/ (2014). Accessed 8 July 2014
47. Thomson, J.J.: Killing, letting die, and the trolley problem. The Monist 59, 204–217 (1976)
48. Transportation Research Board: Animal-vehicle collision data collection. http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_370.pdf (2007). Accessed 8 July 2014
49. Villasenor, J.: Product liability and driverless cars: issues and guiding principles for legislation. Report by The Brookings Institution. http://www.brookings.edu/research/papers/2014/04/products-liability-driverless-cars-villasenor (2014). Accessed 8 July 2014
50. Waterfield, B. and Day, M.: EU has secret plan for police to 'remote stop' cars. The Telegraph. http://www.telegraph.co.uk/news/worldnews/europe/eu/10605328/EU-has-secret-plan-for-police-to-remote-stop-cars.html (2014). Accessed 8 July 2014
51. Wu, S.: Unmanned vehicles and US product liability law. Journal of Law, Information & Science 21(2), 234–254 (2012)
52. Yeomans, G.: Autonomous vehicles: handing over control – opportunities and risk for insurance. Report by Lloyd's. Lloyd's, London. http://www.lloyds.com/~/media/lloyds/reports/emerging%20risk%20reports/autonomous%20vehicles%20final.pdf (2014). Accessed 8 July 2014