J Ethics
DOI 10.1007/s10892-017-9252-2

Incorporating Ethics into Artificial Intelligence

Amitai Etzioni[1] · **Oren Etzioni**[2]

Received: 18 January 2017 / Accepted: 27 February 2017

Abstract This article reviews the reasons scholars hold that driverless cars and many other AI equipped machines must be able to make ethical decisions, and the difficulties this approach faces. It then shows that cars have no moral agency, and that the term ‘autonomous’, commonly applied to these machines, is misleading, and leads to invalid conclusions about the ways these machines can be kept ethical. The article’s most important claim is that a significant part of the challenge posed by AI-equipped machines can be addressed by the kind of ethical choices made by human beings for millennia. Ergo, there is little need to teach machines ethics even if this could be done in the first place. Finally, the article points out that it is a grievous error to draw on extreme outlier scenarios—such as the Trolley narratives—as a basis for conceptualizing the ethical issues at hand.

Keywords Artificial intelligence · Autonomy · Ethics · Self-driving cars · Trolley problem

Driverless cars, which have already travelled several million miles,[1] are equipped with artificial intelligence (AI) that, according to various published claims, enable

[1] Waymo (formerly the Google self-driving car project) alone reports that its driverless cars had logged over two million miles by the end of 2016. See https://waymo.com/journey/.

✉ Amitai Etzioni
etzioni@gwu.edu

Oren Etzioni
orene@allenai.org

[1] The George Washington University, 1922 F Street NW, Room 413, Washington, DC 20052, USA

[2] Allen Institute for Artificial Intelligence, 2157 N. Northlake Way, Suite 110, Seattle, WA 98012, USA

Published online: 07 March 2017

these cars to make autonomous decisions. These decisions have moral and social implications, especially because cars can cause considerable harm. Indeed, in May of 2016, a Tesla car traveling in autopilot mode crashed, and the passenger was killed. (Levin and Woolf 2016) Hence, the question arises: how is one to ensure that the decisions these cars make will be rendered ethically? Speaking more generally, Wendell Wallach and Colin Allen are among those who hold that the world is on the verge of “the creation of (ro)bots whose independence from direct human oversight and whose potential impact on human well-being is the stuff of science fiction.” (Wallach and Allen 2009: 3) That is, the same question asked about driverless cars stands for many other autonomous machines, including weapons that choose their own targets; robotic surgeons; robots that provide child, elder, and health care; as well as quite a few others. This preliminary examination is using driverless cars to examine the ethical questions (encompassing both moral and social values) at hand. (Like many other articles, this article treats the terms *ethical* and *moral* as synonyms).

Specifically, this article first provides a brief overview of the reasoning behind scholars’ assertions that machines guided by AI will need to make ethical decisions; compares two major ways scholars believe this can be achieved—a top-down and a bottom-up approach; asks whether smart machines have or can be given the attributes needed to make moral decisions—to be moral agents; and finally examines to what extent AI-equipped machines are actually autonomous (Sect. 1). In other words, Sect. 1 provides a limited critical review of the relevant literature.

The article then seeks to show that a very significant part of the ethical challenges posed by AI-equipped machines can be addressed by two rather different forms of ethical guidance: law enforcement and personal choices, both used by human beings for millennia. Ergo, there is little need to teach machines ethics even if this could be done in the first place (Sect. 2).

It is crucial for all that follows to note that to the extent that choices are governed by legal dictates— for instance, that a car must stop completely at places marked by a stop sign—the decision of what is ethical has been made *collectively*, through a very familiar process of law- and regulation-making. Here, there is very little need for moral deliberations and decision-making (though there is a need to ‘teach’ the driverless car to comply). As to those decisions left open-ended by law makers—for instance whether the car should stop to give a ride to a hitchhiker—the response is left to each *individual.* For these decisions, the challenge is to find ways for owners or users of driverless cars to guide them in these matters, but not for ethically mandated choices by some collective or third party. True, as we shall see, many decisions involve a mix of collective dictates and individual choices, but each of the two elements of these decisions is still subject to the same considerations when they are faced in a pure form.

1 A Critical Overview

1.1 Reasons Smart Machines are Said to Need Ethics

Driverless cars, viewed as the archetypal autonomous machines, are learning machines. They are programmed to collect information, process it, draw conclusions, and change the ways they conduct themselves accordingly, without human intervention or guidance. Thus, such a car may set out with a program that includes an instruction not to exceed the speed limit, only to learn that other cars exceed these limits and conclude that it can and should speed too. The Tesla car that killed its passenger was traveling at nine miles over the speed limit.

Given that these vehicles may cause harm, scholars argue that driverless cars need to be able to differentiate between ‘wrong’ and ‘right’ decisions. In other words, computers should be made into or become “explicit moral reasoners.” (Wallach and Allen 2009: 6) As Susan Leigh Anderson and Michael Anderson write, “Ideally, we would like to be able to trust autonomous machines to make correct ethical decisions on their own, and this requires that we create an ethic for machines.” (Anderson and Anderson 2011: 1) Many AI researchers seem to hold that if these machines can make thousands of information-driven, cognitive decisions on their own—when to slow down, when to stop, when to yield and so on—they should also be able to make ethical decisions. This assumption is particularly plausible to those who see no fundamental difference between deliberating about factual matters and moral issues, because they view both as mental processes driven by reason.[2] As John Stuart Mill famously wrote, “our moral faculty is a branch of our reason.” (Mill 2008)

Much attention has been paid to the need for these cars (and other AI-equipped, so-called ‘smart’ machines) to choose between two harms in cases when inflicting some harm cannot be avoided. These discussions often begin with an adaptation of the Trolley Problem, wherein the car is unable to brake in time and is forced to choose between continuing in its lane and hitting a pedestrian, or swerving into oncoming traffic in an opposite lane. (Bonnefon et al. 2016) Another variant is that of a child running across the road just before the entrance to a one-lane tunnel, forcing the car to choose between continuing and hitting the child or swerving into the side of the tunnel and killing the passenger. (Millar 2014) In short, driverless cars—and other AI-equipped machines—that make decisions on their own, seem to need ethical guidance.

1.2 Two Ways to Enable ‘Smart’ Cars to Render Ethical Decisions

Two overarching approaches have been suggested as a means of enabling driverless cars and other smart machines to render moral choices on their own: top-down and bottom-up. In the top-down approach, ethical principles are programmed into the

[2] Granted, ‘is’ statements and ‘ought’ statements bleed into each other, but they still differ significantly. Compare a statement against the death penalty that pointed out that data show it does not deter killers, and one that holds that the state should never deliberately take a person’s life. See e.g. McDermott (2011: 88–114).

car’s guidance system. These could be Asimov’s Three Laws of Robotics, the Ten Commandments or other religious precepts—or a general moral philosophy, such as Kant’s categorical imperative, utilitarianism, or another form of consequentialism. The main point is that rather than a programmer instructing the car to proceed under specific conditions in the most ethical way, the car will be able to make such ethical choices based on the moral philosophy that was implanted into its AI program. (Wallach and Allen 2009: 16)

Critics of the top-down approach (as well as some proponents) recognize the inherent difficulties in adhering to any particular moral philosophy, given that any one of them will, at some point or another, lead to actions and outcomes that many will find morally unacceptable. To take but two familiar examples: (1) Benjamin Constant points out that the categorical imperative would obligate someone to tell a murderer the location of his prey, because of the prohibition on lying under any circumstances. (Constant 1797) (2) Obvious concerns would be raised if, following consequentialism, a car concludes that it would be preferable to crash into the less expensive of two cars in adjacent lanes as a way to minimize the amount of damage it causes in a situation where damage is inevitable. (Goodall 2014)

True, these (and other) moral philosophies have developed variants that attempt to address such ‘flaws.’ Still, among and within these schools of ethics, there are significant debates that highlight the difficulties faced in drawing on particular philosophies to serve as moral guidance systems for AI-equipped, smart machines. For instance, there is well-known and significant disagreement over whether and how ‘utility’ can be quantified, with Bentham and Mill disagreeing over whether there are different levels of utility (Mill’s “higher” and “lower” pleasures). Consequentialists continue to face these challenges: for example, estimating long-term consequences and determining for whom consequences should be taken into account. Most of the Trolley Problem thought experiments assume that a body is a body, and hence killing five is obviously worse than one. However, people do not attach the same moral value to terminally ill senior citizens as to children in kindergarten, or to Mother Teresa as to a convicted felon.

There is no need to rehash here the significant back-and-forth between various ethical schools. It suffices to suggest that, given these differences, it is very difficult to program a machine that is able to render moral decisions on its own, whether using one or a combination of these moral philosophies. But one might ask, “If humans can do it, why not smart machines?” In response, one first notes that humans are able to cope with nuance and deal with fuzzy decisions, but computer programmers find such decisions particularly taxing. Moreover, while one can argue that individuals make moral choices on the basis of this or that philosophy, actual humans first acquire moral values from those who raise them, and then modify these values as they are exposed to various inputs from new groups, cultures, and subcultures, gradually developing their own personal moral mix. Moreover, these values are influenced by particular societal principles that are not confined to any one moral philosophy. In short, the top-down approach is highly implausible.

In the second approach to machine ethics, the bottom-up approach, machines are expected to learn how to render ethical decisions through observation of human behavior in actual situations, without being taught any formal rules or being equipped with any particular moral philosophy. This approach has been applied to non-ethical aspects of driverless cars' learning. For example, an early autonomous vehicle created by researchers at Carnegie Mellon University was able to navigate on the highway after 2–3 min of training from a human driver; its capacity for generalization allowed it to drive on four-lane roads, even though it was only trained on one- or two-lane roads. (Batavia et al. 1996) Machine learning has also been used by several researchers to improve a car's pedestrian detection ability.[3]And, a team from NVIDIA Corporation recently demonstrated a driverless car that used 'end-to-end' machine learning, which was able to drive on its own after observing only 72 h of human driving data. (Bojarski et al. 2016)

However, to view these as precedents for learning ethical conduct is to presume that there is no significant difference between learning to respond differently, say, to green, red, and yellow traffic lights, and—learning to understand and appreciate the moral imperative to take special care not to hit a bicyclist traveling in the same lane as the car, let alone not to harass or deliberately hit the cyclist out of road rage. (McDermott 2011) However, the kinds of moral questions the cars are asked to address—who to kill or injure in a situation where a crash is inevitable—are actually very rare. According to data from the US Department of Transportation, there were only 77 injuries and 1.09 fatalities per 100 million miles driven in 2013. (National Highway Traffic Safety Administration 2013) And, each such challenging situation is different from the next one: sometimes it is a kitten that causes the accident, sometimes a school bus, and so on. A driverless car would have to follow a person for several lifetimes to learn ethics in this way.

It has hence been suggested that driverless cars could learn from the ethical decisions of millions of human drivers, through some kind of aggregation system, as a sort of groupthink or drawing on the wisdom of the crowds. One should note, however, that this may well lead cars to acquire some rather unethical preferences, as it is far from clear that the majority of drivers would set a standard worthy of emulation by the new autonomous cars. If they learn what many people do, smart cars may well speed, tailgate, and engage in road rage. One must also note that people may draw on automatic reflexes when faced with the kind of choices posed by the Trolley Problem rather than on ethical deliberations and decision-making. That is, observing people will not teach these machines what is ethical, but what is common.

This concern is supported by an experiment conducted by Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan, who tested participants' attitudes about whether driverless cars should make utilitarian moral decisions, even when that would mean sacrificing the passenger's life in order to save a greater number of pedestrians. They found that most respondents want driverless cars to make utilitarian decisions as long as they are not involved; for themselves they desired cars that will prioritize their own well-being at the cost of others. (Bonnefon et al. 2016) As philosopher Patrick Lin put it,

[3] See Hsu (2016) and Harris (2015).

"No one wants a car that looks after the greater good. They want a car that looks after them." (Metz 2016) This is hardly a way for Google, Tesla, or any other car manufacturer to program ethical cars. They best not heed the voice of the masses.

In short, both the top-down and the bottom-up approaches face very serious difficulties. These difficulties are not technological but concern the inner structures of ethical philosophies used by humans. Even so, these difficulties pale in comparison to those posed by the question of whether or not smart machines can be turned into moral agents in the first place.

1.3 How Autonomous are Smart Machines?

In many discussions of the ethical challenges posed by driverless cars, and smart machines generally, such machines are referred to as 'autonomous.' To begin with, one must recall that not every scholar is willing to take it for granted that even human beings act autonomously. Some hold that everything that happens is caused by sufficient antecedent conditions which make it impossible for said thing to happen differently (or to not happen); such causal determinism renders it impossible to assign moral responsibility.[4] There is no need here to repeat the arguments against this position; it suffices to note that we file with those who take it for granted that human beings have some measure of free will, though much of their lives may indeed be determined by forces beyond their understanding and control. (Frankfurt 1969)

However, it does not follow that the same holds for machines, however smart they are. Indeed, a colleague who read a previous draft of this article argued that it only *seems* like smart machines make decisions on their own—in actuality, changes in how these machines conduct themselves merely reflect external forces. One could say, he pointed out, that a missile was diverted from its original course because a strong gust of wind 'decided' to change direction, but this would be merely a misperception or illusion.

As we see it, autonomy is a variable that exists along a continuum. Some tools have no autonomy—one can fully account for their actions by forces external to them. A hammer hitting a nail has no autonomy even when it misses, because one can show that the miss was due to inexperience of the person using it, poor eyesight, or some other external factor. A rudimentary GPS system may be said to have a very small measure of autonomy because, when asked the best way to get from point A to point B, it compares several options and recommends one, but its recommendation is based on a human-made algorithm that calculates the shortest route, or that which will take the least amount of time to travel, or some other implanted criteria. A significant amount of autonomy occurs when the machine is given a large number of guidelines, some that conflict with each other, and is ordered to draw on information it acquires as it proceeds, and to draw conclusions on its own—such as a more advanced GPS system, which identifies upcoming traffic, or an accident, and

[4] See e.g. van Inwagen (1997: 373–381) and Harris (2011).

reroutes accordingly. Machines equipped with AI are held to be able to act much more autonomously than those not so equipped.

Monica Rozenfield writes:

> Deep learning is a relatively new form of artificial intelligence that gives an old technology—a neural network—a twist made possible by big data, supercomputing, and advanced algorithms. Data lines possessed by each neuron of the network communicate with one another.
>
> It would be impossible to write code for an unlimited number of situations. And without correct code, a machine would not know what to do. With deep learning, however, the system is able to figure things out on its own. The technique lets the network form neural relationships most relevant to each new situation. (Rozenfield 2016)

A group of computer scientists from Carnegie Mellon University notes that "Machine-learning algorithms increasingly make decisions about credit, medical diagnoses, personalized recommendations, advertising and job opportunities, among other things, but exactly how usually remains a mystery." (Spice 2016)

Some believe that machines can attain full autonomy: for instance, weapon systems that choose their own targets without human intervention, and whose missions cannot be aborted. In fact, even these machines are limited to the missions set for them by a human, and they are only 'free' to choose their targets because a human programmed them that way. Their autonomy is limited.

Military ethicist George Lucas, Jr. notes that debates about machine ethics are often obfuscated by the confusion of machine autonomy with moral autonomy; the Roomba vacuum cleaner and Patriot missile are both autonomous in the sense that they perform their missions, adapting and responding to unforeseen circumstances with minimal human oversight—but not in the sense that they can change or abort their mission if they have moral objections. (Lucas Jr. 2013) Pedro Domingos writes:

> They can vary what they do, even come up with surprising plans, but only in service of the goals we set them. A robot whose programmed goal is "make a good dinner" may decide to cook a steak, a bouillabaisse, or even a delicious new dish of its own creation, but it can't decide to murder its owner any more than a car can decide to fly away. (Domingos 2015: 283)

Brad Templeton put it well when he stated that a robot would be truly autonomous the day it is instructed to go to work and it instead goes to the beach. (Markoff 2015: 333)

For the sake of the following discussion, we shall assume that smart machines have a significant measure of autonomy, and surely a greater capacity to render cognitive choices on their own than old fashioned machines (e.g. the ability to decide on their own how much to slow down when the roads are slick, without a programmed instruction that covers such a condition). Given this measure of autonomous volition, these cars are potentially more likely to be able to choose to cause harm, and therefore require ethical guidance, but they do not necessarily have an ability to make ethical choices autonomously, as we shall see. Machines are ultimately tools of the human beings who design and manufacture them.

1.4 When Smart Machines Stray

So far we have referred to smart machines as many AI scholars do—as autonomous machines. However, 'autonomous' is a highly loaded term because in liberal democracies it is associated with liberty, self-government, and individual rights. To violate someone's autonomy is considered a serious ethical offense (although one acknowledges that there are some extenuating circumstances). Indeed, bioethicists consider autonomy to be the leading most important principle guiding patient care: physicians and other health care personnel should first and foremost heed the preferences of the patient. However, cars and other machines are not emotional beings that experience pain or shame, but unfeeling tools made to serve humans. *There is nothing morally objectionable about overriding their choices, or making them toe the line*. One does not violate their dignity by forcing them to make choices within the boundaries set by their programmers. While we would be horrified if one rewired the brain of an autonomous person, there is no ethical reason to object to reprogramming a smart machine that is causing harm to human beings.

A basic change in the way these machines are conceptualized serves to highlight our point: if a car that decided on its own to speed or tailgate was considered a rule-breaking offender or a deviant (i.e. an agent that deviated from the prevailing norms), one would ask how to reprogram that car in order for it to "behave" better. One would not ask—as one does about an autonomous person—how to help that car acquire the moral values that would allow it to make more appropriate ethical decisions. (How machines can be reined in is discussed below). To push the point: human beings—even if they have a highly developed sense of right and wrong and score highly on the various attributes that make them moral agents—occasionally misbehave. And when they do, society tries to draw out their good nature and improve their character by moral suasion and re-education—but often, society will also set new limits on them (curfew for teenagers and jail for repeat drug offenders). There seems no reason to treat cars any differently. Indeed, since a malfunctioning smart car is not 'autonomous' in the way that people are, there appear to be no moral injunctions against implementing extensive constraints on a smart car's actions. Quite simply, the car is a malfunctioning tool that should be dealt with accordingly.

1.5 Partners: Not Free Standing Agents

Another major source of the misconception that seems to underlie much of the discussion is found in public discussions of AI and even some academic ones, namely the assumption that there is essentially one kind of program that makes machines much more effective and efficient ('smarter')—a guidance system that draws on artificial intelligence. Actually, there are two different kinds of AI. The first kind of AI involves software that seeks to reason and form cognitive decisions the way people do (if not better), and thus aspires to be able to replace humans. It seeks to reproduce in the digital realm the processes in which human brains engage when they deliberate and render decisions. The famous Turing test deals with this kind of AI; it deems that a program qualifies as ‘intelligent’ if its reactions are indistinguishable from that of a person. One could call this kind of AI *AI minds*.

The other kind of AI merely seeks to provide smart assistance to human actors—call it *AI partners*. This kind of AI only requires that the machines be better at rendering decisions in some matters than humans, and that they do so effectively within parameters set by humans or under their close supervision. For instance, AI caregivers engage in childcare in conjunction with parents, taking care of the children for short periods of time, or while parents are working nearby within the home. (Sharkey and Sharkey 2010) When AI caregivers engage in elder care, they work in conjunction with human personnel, carrying out some tasks on their own (e.g. reminding patients of the time to take medication and chatting with them when they are lonely) but alerting the human staff in response to certain conditions (e.g. a patient leaving the room). (Sharkey and Sharkey 2012)

Those who seek to call attention to the key difference under discussion have used a wide variety of other terms. Terms used in reference to AI partners include “Intelligence Augmentation” (IA) (Markoff 2015), “intelligence amplification,” “cognitive augmentation,” and “machine augmented intelligence.” (DMello 2015) (John Markoff dedicates much of his book *Machines of Loving Grace: The Quest for Common Ground Between Humans and Robots,* to the difference between these two camps, their major figures, and the relations between them). Some AI mavens hold that the reason they pay little attention to the difference between the two AIs is that the work they do applies equally to both kinds of AI. However, often—at least in public discourse—the difference is significant. For instance, the threat that AI will make machines so smart that they could dominate humans[5] applies mainly to AI minds but not AI partners.

In terms of cars, Google is developing a completely driverless car, going so far as to remove the steering wheel and brake pedal from its recent models; this is an example of AI minds. Tesla merely seeks (at least initially) to provide human drivers with AI features that make driving safer. Passengers are warned that even when the car is in autopilot mode, they must keep their hands on the wheel at all times and be an alert partner driver. True, as AI partners become more advanced, the difference between the two kinds of AI could shrink and one day disappear. For now, the opposite problem prevails: namely, that AI partners with rather limited capabilities are expected to act (or evoke fears) as if they were AI minds.

All of this is pertinent because if smart machines are going to have minds, replace humans, and act truly on their own—and if humans are to be removed from the loop (e.g. in killing machines that cannot be recalled or retargeted once they are launched)—then smart machines will indeed have to be able to render moral decisions on their own. As their volition increases, smart machines will have to be treated as if they were moral agents and assume at least some responsibility for their acts. One could no longer consider only the programmers, manufactures, and

[5] This is, of course, a popular theme in science fiction, but for a serious treatment of the threat, see Joy (2000).

owners (from here on, the term also refers to users) as the moral agents. Under this condition, the question of who or what to ticket when a driverless car speeds becomes an acute one.

However, there seem to be very strong reasons to treat smart machines as partners, rather than as commanding a mind that allows them to function on their own. A major reason is that although smart machines seem very good at carrying out some functions that humans used to perform (e.g. memorizing), they are very poor at others (e.g. caring about those they serve and others). Thus, elder care robots are good at reminding patients to take their medications, but not at comforting them when they grieve or are fearful, as compared to two-legged mortals.

In particular, at least for the foreseeable future, a division of labor between smart machines and their human partners calls for the latter to act as the moral agent. Human beings have the basic attributes needed for moral agency, attributes that smart machines do not have and which are very difficult to implant into them. The article turns next to examine how humans can provide moral guidance to smart machines, despite the fact that they are learning machines and hence have a strong tendency to stray from the instructions originally programmed into them.

2 The Main Factors for Implementing Ethics

2.1 Legal and Personal

How can driverless cars and other such machines follow the ethical preferences of humans if these machines do not have a capacity to make ethical decisions on their own? In answering this question, one must first consider the two very different ways that moral and social values are implemented in the human world, and then how these values might be introduced into the realm of smart machines.

The primary ways moral and social values are implemented in society are through legal enforcement and personal choices (although these choices are socially fostered). Many moral and social values are embodied in laws (and their expression in regulations), i.e. they are collectively formulated and enforced. For example, in order to enhance safety, a car must come to a full stop at stop signs. Those who do not heed these values are physically prevented from continuing (e.g. their cars are towed if they park illegally in a fire lane), penalized (e.g. issued tickets), or jailed (e.g. drunken drivers). In sharp contrast, heeding other values is left to each individual's choices, based on their moral preferences (e.g. stopping to help stranded motorists). Compliance with these values is fostered through informal social controls. Those who violate these values may be shamed or chastised, while those who abide by them are commended and appreciated. But the final decision on matters such as whether or not one buys environmentally friendly gasoline or the cheapest available, purchases a car that pollutes less than others, stops for hitchhikers, or allows friends to use one's car is left to each individual.

The distinction between the two modes of implementing social and moral values—between legal enforcement and personal choices—is critical because the *many values* that are implemented through laws enacted by legislatures, interpreted by courts, and enforced by the state *are in principle not subject to individual deliberations and choice.* They are subject to collective deliberation and decisions. Society does not leave it to each individual to decide if he or she holds that it is morally appropriate not to speed, nor tailgate, pass only on the left (usually), refrain from running through stoplights, pollute, throw trash out of the window, wear a seat belt, leave the scene of a crash, drive intoxicated, or drive under the legal age, among many other decisions. Hence, the notion that smart machines need to be able to render moral decisions does not take into account that in many of these important matters, *what cars ought to do is not up to them any more than it is up to their owners or users*.

By treating all these choices as 'ethical' in nature (which they are), but disregarding the many facets of human behavior and decision-making that are not subject to individual deliberation and decision, the advocates of machine ethics see a much greater realm of ethical decision-making for the AI-equipped machine than actually exists. Driverless cars will have to obey the law like all other cars and there is no need for them to be able to deliberate if they consider it ethical to speed, pollute, and so on. True, these laws may be adapted to take into account the special features of these cars (e.g. allowing them to proceed at a higher speed than other cars in their own lane.) Still, driverless cars will need to obey the laws, collectively agreed upon, like all other cars, or else be taken off the road. Their owners, programmers, and manufacturers will need to be held liable for any harm done.[6]

The ethical decisions that are not prescribed by law are left to be made by individuals or—their cars. We already have seen that seeking to program these cars to be able to make these decisions on their own is, at best, a very difficult task. What can be done? One answer is for individuals to instruct the car they own or use to follow their value preferences. To a limited extent, this can be achieved through setting options. For instance, Tesla enables owners to set the distance their car maintains from the car in front of it. (Gibbs 2015) However, data show that people tend not to engage in such decision-making if they must make more than a few choices. Numerous studies of human behavior, ranging from retirement contributions[7] to organ donations (Johnson and Goldstein 2003) to consumer technology (Shah and Sandvig 2008) reveal that the majority of people will simply choose the default setting, even if the options available to them are straightforward and binary (e.g. opt-in vs. opt-out). This is not to suggest that customization should be excluded but to acknowledge that it cannot take care of most of the personal choices that must be made.

To proceed, we suggest that enhanced moral guidance to smart machines should draw on a new AI program that will 'read' the owner's moral preferences and then instruct these machines to heed them. We call it an *ethics bot*. An ethics bot is an AI program that analyzes many thousands of items of information (not only information publicly available on the Internet but also information gleaned from a person's local computer storage and that of other devices) about the acts of a particular individual in order to determine that person's moral preferences.

[6] See Etzioni and Etzioni (2016a: 149–156, 2016b: 133–146).

[7] See Beshears et al. (2009: 167–195) and Benartzi and Thaler (2013: 1152–1153).

Essentially, what ethics bots do for moral choices is similar to what many AI programs do when they are ferreting out consumer preferences and targeting advertising to them accordingly. For instance, an ethics bot may conclude that a person places high value on environmental protection if it finds that said person purchases recycled paper, drives a Prius, contributes to the Sierra Club, prefers local food, and never buys Styrofoam cups. It would then instruct that person's driverless car to refuel using only environmentally friendly gas, to turn on the air conditioning only if the temperature is high, and to turn off the engine at stops. One should note that ethics bots do not seek to teach the car an ethical philosophy that will enable it (or other smart machines) to deliberate and then form its own moral conclusions. Rather, the ethics bot extracts specific ethical preferences from a user and subsequently applies these preferences to the operations of the user's machine.

To illustrate: Nest constructed a very simple ethics bot, which has already been used by more than a million people. Nest built a smart thermostat which first 'observes' the behavior of the people in their households for a week, noting their preferences on how cool or warm they want their home to be. The smart thermostat then uses a motion-detecting sensor to determine whether anyone is at home. When the house is empty, the smart thermostat enters into a high energy saving mode; when people are at home, the thermostat adjusts the temperature to fit their preferences. This thermostat clearly meets the two requirements of an ethics bot, albeit a very simple one: it assesses people's preferences and imposes them on the controls of the heating and cooling system. One may ask what this has to do with social and moral values. This thermostat enables people with differing values to have the temperature settings they prefer—to be either more environmentally conscious or less so. (Lohr 2015)

For such an approach to be applied to more complex ethical issues, considerable additional work must be done. One notes that even programs that seek to ferret out consumer preferences—for instance, to suggest movies or books that consumers may like—are still works in progress. Ethics bots are hence best viewed as a research agenda rather than programs that one can take off the shelf.

One may say that ethics bots are very much like the bottom-up approach we viewed as visionary. However, the ethics bot approach does not require that the machines learn to adopt any kind of ethics or have any of the attributes of moral agents. The ethics bot merely 'reads' the specific moral preferences of the user and instructs the machine to heed them. One may ask: what if these preferences are harmful? Say the ethics bot orders the car to speed in a school zone because that is what the owner would do. This question and similar ones do not take into account the major point we cannot stress enough: that the ethical decisions left to the individual are only those which the society ruled—rightly or wrongly—are not significantly harmful, and hence remain unconstrained by regulation or attendant legislation. That is, individuals are free to make them as they see fit. (Moreover, societies constantly change the extent and kind of behavior they regulate *collectively*—through changes in law and levels of enforcement).

We earlier argued that if cars learn from individuals how to proceed, they may learn to engage in 'bad' behavior. This is the case if the smart machines are to gain all of their moral guidance from following what individuals do. In contrast, ethics bots only address areas left opened-ended by the law.

So far, as a first approximation, we discussed choices as if they were either legal, and hence collective, or ethical, and hence personal. However, a fair number of decisions involve elements of both. For instance, the requirement for cars to yield to pedestrians under many circumstances is prescribed by law. Still, people who have a strong sense of respect for others are more likely to take extra precautions not to even come near pedestrians than those who have a weaker commitment to this value. Hence, cars will need to be guided by ethics bots to determine whether they are to get close to the line defined by law or to give it a wide margin.

Finally, one must take into account that decisions are affected by many considerations other than the adherence to moral and social values—for instance, by interpretations of reality and by emotions. A comparison of human-driven cars and driverless cars that would encompass all these factors is likely to find that for many of them, the driverless cars will score much better than the human-driven ones. This is especially likely as the technologies involved are further developed. This observation has a major ethical implication as one must expect that in the foreseeable future, driverless cars will become much safer than human-driven ones.

To review the discussion so far: We have seen that implanting ethics into machines, or teaching machines ethics is, at best, a very taxing undertaking. We pointed out that many of the ethical decisions that smart machines are said to have to make, need not and should not be made by them because they are entrenched in law. These choices are made for the machines by society, using legislatures and courts. Many of the remaining ethical decisions can be made by ethics bots, which align the cars' 'conduct' with the moral preferences of the owners. Granted, neither the law nor ethics bots can cover extreme outlier situations. These are discussed next.

2.2 The Outlier Fallacy

A surprising amount of attention has been paid to the application of Trolley Problems to driverless cars. The media frequently uses these tales as a way to frame the discussion of the issues at hand, as do a fair number of scholars. The Trolley Problems are not without merit. Like other mental experiments, they can serve as a provocative dialogue starter. And they can be used as an effective didactic tool, for instance to illustrate the difference between consequentialism and deontology. However, such tales are particularly counterproductive as a model for decision-making by smart machines and their human partners. The Trolley Problems are extremely contrived. They typically leave the actor with only two options; neither of these options nor any of the other conditions can be modified. For example, the choice is framed as either killing a child or causing a devastating pile up. To further simplify the scenario, it assumes that killing two people is 'obviously' worse than one, disregarding that most people value different people's lives very differently; compare a 95-year-old person with terminal cancer to a 25-year-old war veteran, or to a child. James O'Connor adds:

> What is wrong with trolley theorizing is that by design it implicitly, but nonetheless with unmistakable dogmatism, stipulates that the rescuer is not in a position, or does not have the disposition, to really help, only to act by selecting one or other of a Spartan range of choices, all of them morally repugnant, that the trolley philosopher has pre-programmed into the scenario. The trolley method, by this token, is premised on a highly impoverished view of human nature. (O'Connor 2012: 245)

Barbara Fried suggests that the "intellectual hegemony" of trolley-ology has encouraged some philosophers to focus more on "an oddball set of cases at the margins" than on the majority of real-life cases where the risk of accidental harm to others actually occurs. (Fried 2012)

An important adage in legal scholarship is that "hard cases make bad law;" cases that attract attention because of particularly extreme circumstances tend to result in laws or decisions that address the exception but make for poor rules. The same holds for ethics. Thus, the "Ticking Time Bomb" scenario is used to argue that utilitarian ethics justifies torture. (Luban 2005) And, just because some may prostitute themselves if promised that in exchange, their spouse's life will be spared, that does not mean that everyone has a price, or that everyone is willing to prostitute themselves. (Wood 2007)

We observe that this a broad issue for statistical machine learning programs (and for statistical inference more broadly). When generalizing from a sample, drawn from some probability distribution, there are outlier events (e.g. 4 or 5 standard deviations from the mean of a normal distribution) where it is difficult to predict the outcome. Several responses are in order. First, by definition, these are 'outliers'—rare occurrences whose frequency is bounded. Second, human drivers face the same issues and often react in an unpredictable fashion, which sometimes results in so-called 'human error.' Third, advanced machine learning systems generalize from previous experience and do cover never-before-seen situations regularly. Overall, this remains a challenging research topic for machine learning, but one that is not unique to this arena.

To reiterate, most of the time, smart machines can be kept in line through legal means. In other situations, they can be made to abide by their users' ethical preferences. (In both cases they are assisted by second order, supervisory AI programs.) Granted, these two moral guidance modalities will leave uncovered the once-in-a million-miles situations (each unique and often unpredictable). There will always be some incidents that cannot be foreseen and programmed; this happened, for instance, when a metallic balloon flew right in front of a driverless car and confused it. In these cases, the choice is best left to be made randomly with regards to which party will be harmed (covered by no-fault insurance). If these situations are repeated, the legal guidance or that of the ethics bot will need to be updated by humans.

When all is said and done, it seems that one need not, and most likely cannot, implant ethics into machines, nor can machines pick up ethics as children do, such that they are able to render moral decisions on their own. Society can set legal limits on what these machines do in many cases; their users can provide them with ethical guidance in other situations, employing ethics bots to keep AI-equipped machines in line; and one can leave alone the one-in-a-million situations, without neglecting to compensate those that are harmed.

References

Anderson, Michael, and Susan Leigh Anderson (eds.). 2011. *Machine ethics*. Cambridge: Cambridge University Press.

Batavia, Parag H., Dean A. Pomerleau, and Charles E. Thorpe. 1996. Applying advanced learning algorithms to ALVINN. Carnegie Mellon University, The Robotics Institute. http://www.ri.cmu.edu/pub_files/pub1/batavia_parag_1996_1/batavia_parag_1996_1.pdf.

Benartzi, Shlomo, and Richard H. Thaler. 2013. Behavioral economics and the retirement savings crisis. *Science* 339: 1152–1153.

Beshears, John, James J. Choi, David Laibson, and Brigitte C. Madrian. 2009. The importance of default options for retirement saving outcomes: evidence from the United States. In *Social security policy in a changing environment*, ed. David A. Wise, and Jeffrey B. Liebman, 167–195. Chicago: University of Chicago Press.

Bojarski, Mariusz, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel et al. 2016. End to end learning for self-driving cars. arXiv. https://arxiv.org/abs/1604.07316.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352: 1573–1576.

Constant, Benjamin. 1797. Des réactions politiques. *Oeuvres complètes* 1: 1774–1799.

DMello, Alvin. 2015. Rise of the humans: intelligence amplification will make us as smart as the machines. The Conversation. http://theconversation.com/rise-of-the-humans-intelligence-amplification-will-make-us-as-smart-as-the-machines-44767.

Domingos, Pedro. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.

Etzioni, Amitai, and Oren Etzioni. 2016a. AI assisted ethics. *Ethics and Information Technology* 18: 149–156.

Etzioni, Amitai, and Oren Etzioni. 2016b. Keeping AI legal. *Vanderbilt Journal of Entertainment and Technology Law* 19: 133–146.

Frankfurt, Harry G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66: 829–839.

Fried, Barbara H. 2012. What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly* 62: 505–529.

Gibbs, Samuel. 2015. What's it like to drive with Tesla's autopilot and how does it work? The Guardian. https://www.theguardian.com/technology/2016/jul/01/tesla-autopilot-model-s-crash-how-does-it-work.

Goodall, Noah. 2014. Ethical decision making during automated vehicle crashes. *Transportation research record: journal of the transportation research board* 2424: 58–65.

Harris, Mark. 2015. New pedestrian detector from Google could make self-driving cars cheaper. IEEE Spectrum. http://spectrum.ieee.org/cars-that-think/transportation/self-driving/new-pedestrian-detector-from-google-could-make-selfdriving-cars-cheaper.

Harris, Sam. 2011. *The moral landscape: How science can determine human values*. New York: Simon and Schuster.

Hsu, Jeremy. 2016. Deep learning makes driverless cars better at spotting pedestrians. IEEE Spectrum. http://spectrum.ieee.org/cars-that-think/transportation/advanced-cars/deep-learning-makes-driverless-cars-better-at-spotting-pedestrians.

Johnson, Eric J., and Daniel Goldstein. 2003. Do defaults save lives? *Science* 302: 1338–1339.

Joy, Bill. 2000. Why the future doesn't need us. WIRED. http://www.wired.com/2000/04/joy-2/.

Levin, Sam and Nicky Woolf. 2016. Tesla driver killed while using autopilot was watching Harry Potter, witness says. The Guardian. https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter.

Lohr, Steve. 2015. Homes try to reach smart switch. *New York Times*. http://www.nytimes.com/2015/04/23/business/energy-environment/homes-try-to-reach-smart-switch.html?_r=0.

Luban, David. 2005. Liberalism, torture, and the ticking bomb. *Virginia Law Review* 91: 1425–1461.

Lucas Jr., George R. 2013. Engineering, ethics and industry: the moral challenges of lethal autonomy. In *Killing by remote control: the ethics of an unmanned military*, ed. Bradley Jay Strawser, 211–228. New York: Oxford University Press.

Markoff, John. 2015. *Machines of loving grace: the quest for common ground between humans and robots*. New York: ECCO.

McDermott, Drew. 2011. What matters to a machine. In *Machine ethics*, ed. Michael Anderson, and Susan Leigh Anderson, 88–114. Cambridge: Cambridge University Press.

Metz, Cade. 2016. Self-driving cars will teach themselves to save lives—but also take them. WIRED. http://www.wired.com/2016/06/self-driving-cars-will-power-kill-wont-conscience/.

Mill, John Stuart. 2008. *On liberty and other essays*. Oxford: Oxford University Press.

Millar, Jason. 2014. You should have a say in your robot car's code of ethics. WIRED. http://www.wired.com/2014/09/set-the-ethics-robot-car/.

National Highway Traffic Safety Administration. 2013. Traffic safety facts 2013: A compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. US Department of Transportation. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812139.

O'Connor, James. 2012. The trolley method of moral philosophy. *Essays in Philosophy* 13: 242–255.

Rozenfield, Monica. 2016. The next step for artificial intelligence is machines that get smarter on their own. The Institute. http://theinstitute.ieee.org/technology-topics/artificial-intelligence/the-next-step-for-artificial-intelligence-is-machines-that-get-smarter-on-their-own.

Shah, Rajiv C., and Christian Sandvig. 2008. Software defaults as de facto regulation the case of the wireless Internet. *Information, Community and Society* 11: 25–46.

Sharkey, Amanda, and Noel Sharkey. 2012. Granny and the robots: ethical issues in robot care for the elderly. *Ethics and Information Technology* 14: 27–40.

Sharkey, Noel, and Amanda Sharkey. 2010. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies* 11: 161–190.

Spice, Byron. 2016. Carnegie Mellon transparency reports make AI decision-making accountable. Carnegie Mellon Computer University School of Computer Science. http://www.cs.cmu.edu/news/carnegie-mellon-transparency-reports-make-ai-decision-making-accountable.

Van Inwagen, Peter. 1997. Fischer on moral responsibility. *The Philosophical Quarterly* 47: 373–381.

Wallach, Wendell, and Colin Allen. 2009. *Moral machines: teaching robots right from wrong*. New York: Oxford University Press.

Waymo, https://waymo.com/journey/. Accessed 14 Feb 2017.

Wood, Allen. 2007. *Kantian ethics*. Cambridge: Cambridge University Press.