Chapter 15
Incorporating Ethics into Artificial Intelligence (with Oren Etzioni)

Driverless cars, which have already travelled several million miles,[1] are equipped with artificial intelligence (AI) that, according to reports, enable these cars to make autonomous decisions. These decisions have moral and social implications, especially because cars may 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). Wallach and Allen (2009, p. 3) are among those who hold that the world is on the verge of "the creation of robots whose independence from direct human oversight and whose potential impact on human well-being is the stuff of science fiction." Hence, the question arises: how is one to ensure that the decisions of these cars will be rendered ethically? The same question stands for other autonomous machines: weapons that choose their own targets; robotic surgeons; robots that provide child, elder, and health care; as well as quite a few others.

Several scholars have addressed this new challenge by suggesting that driverless cars (and other autonomous machines) be programmed to be able to render moral decisions on their own. (Like many other articles, this chapter treats the terms *ethical* and *moral* as synonyms.) This chapter attempts to show that such a course is, at best, enormously taxing; the notion that autonomous machines can be made to render ethical decisions is based on a conflation of several misconceptions, including what criteria are required of a moral agent, and about the fundamental nature of human decision-making. This line of criticism, in turn, leads to a suggestion that we require a much more 'traditional' (i.e. in line with the world before machines became autonomous) solution to the ethical challenges posed by these machines.

Specifically, this chapter (a) provides a brief overview of the reasons that scholars hold that these machines will be able to make ethical decisions, and enumerates the very considerable difficulties this approach faces. (b) It shows that the way the

---

This chapter draws on "Incorporating Ethics into Artificial Intelligence" in *The Journal of Ethics*, 21 (4), (2017): 403–418.

[1] As of July 2016 Google alone reports that its driverless cars have logged 1.5 million miles. See https://www.google.com/selfdrivingcar/

A. Etzioni, *Happiness is the Wrong Metric*, Library of Public Policy and Public Administration 11, https://doi.org/10.1007/978-3-319-69623-2_15

term "autonomous" is commonly applied to these machines is misleading, and that it leads to invalid conclusions about the kinds of AI used in these machines. The chapter's most important claim is that (c) a very significant part of the challenge posed by these 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. (d) The chapter then points out that it is a grievous error to draw on extreme outlier scenarios as a basis for conceptualizing the issues at hand, and that one reaches radically different conclusions once the extreme nature of these examples (associated with the Trolley Problem) is recognized.

15.1 Smart Machines, Harm, and Ethical Self-Guidance

15.1.1 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 their behavior 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.

Given that 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, p. 6). Susan Leigh Anderson and Michael Anderson (2011, p. 1) argue, "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." 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 (1859/2008) famously wrote, "our moral faculty is a branch of our reason."

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 of unavoidable harm. 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 continu-

[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).

ing 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).

15.1.2 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 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 it was granted or acquired (Wallach and Allen 2009, p. 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 some will find morally unacceptable. To take but two familiar examples: 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). As for consequentialism, if a car's instruction is to minimize the amount of damage it causes in a situation where damage is inevitable, it may calculate that it would be preferable to crash into the less expensive of two cars in the adjacent lanes, which would raise obvious concerns about discrimination (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 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 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 among 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, deal with fuzzy decisions while 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, actually 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 (See Hsu 2016; Harris 2015). 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 to green, red, and yellow traffic lights and learning to understand and appreciate the 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). But this parallel between cognitive and ethical decision-making is far from self-evident. Some AI mavens believe that cars could learn ethics like children do, gradually improving their ability to engage in moral reasoning, along the Kohlberg stages. We shall see that cars may be able to follow specific moral positions, but not necessarily to engage in moral reasoning.

Moreover, 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: 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 group think or drawing on the wisdom of the crowds. 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 responses 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 et al. (2016), 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, but they themselves desire cars that will prioritize their own well-being at the cost of others. Philosopher Patrick Lin put it that, "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'd best not heed the masses' voice.

In short, both the top-down and the bottom-up approaches face very serious difficulties. These difficulties are not of a mechanical sort, 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.

15.1.3 Can Smart Machines Be Made into Moral Agents?

Both the top-down and bottom-up approaches to machine ethics presume that machines can be made into or become moral agents. Moral agency is understood to be the ability to make decisions based on some conception of morality, and to act in a manner that can be judged as either "right" or "wrong." (If a printer malfunctions, for example, one does not consider it to have done something "wrong.") To be a moral agent requires a specific set of attributes. Drawing on the work of Eugene Hargrove (1992) and of Luciano Floridi and J.W. Sanders (2004), we suggest that at least the following attributes are essential: (a) Self-consciousness. If the agent is not aware of itself in any given situation, and of the alternative courses that might be followed, then no moral decisions can be rendered. (b) The agent must be aware that she can affect the situation. (c) The agent must be able to understand the moral principles to be employed in arriving at a particular moral choice. (d) The agent must have a motive to act morally. This involves having passions, as otherwise moral preferences are merely intellectual preferences with nothing to fuel the moral

action.[3] (e) Some scholars add that a will or intention is required (see Coeckelbergh 2009; Himma 2009).

One notes that autonomous machines do not have the attributes required for them to be able to act as moral agents. We are hardly the first or the only ones to make this crucial observation. Domingos (2015, p. 283) writes "Unlike humans, computers don't have a will of their own. They're products of engineering, not evolution. Even an infinitely powerful computer would still be only an extension of our will." Deborah Johnson (2006) argues that while computer systems can in fact have intentionality and thus should not be dismissed from the realm of morality, because they do not possess mental states, they cannot be considered moral agents. Patrick Chisan Hew (2014) holds that a central issue in establishing moral agency is responsibility, but any machine will have its rules for behavior supplied externally by humans, thus denying it moral agency.

Moreover, intelligence and intentionality may go together in humans, but cars can only have the first, not the second. Pedro Domingos (2015) observes:

> It's natural to worry about intelligent machines taking over because the only intelligent entities we know are humans and other animals, and they definitely have a will of their own. But there is no necessary connection between intelligence and autonomous will; or rather intelligence and will may not inhabit the same body, provided there is a line of control between them.

Some scholars tried to deal with this challenge by arguing that although cars cannot be made into "full ethical agents," they can be "ethical impact agents" or "implicit ethical agents" (Moor 2011). The first is defined as agents that have a moral impact, even if unintended and not a result of any moral deliberation. By this definition, falling rocks, runaway fires, and hurricanes are ethical impact agents. The second is defined as agents that are designed with some sort of virtue in mind; for example, ATM machines that protect privacy. But of course, the ATM's virtues are fully implanted by humans. These two definitions seem to stretch exceedingly the concept of moral agency because they do not entail machines that are engaging in moral deliberations and acting on their conclusions.

Other valiant efforts to support machine ethics leads Susan Leigh Anderson (2011) to argue that cars' lack of emotions (and therefore empathy) would not necessarily be a weakness, as humans often get "carried away" by their emotions and thus behave immorally or unethically. Others go much further. Joseph Emilie Nadeau (2006) contends that an action can only be said to have been taken freely if it is the product of a strictly logical, fully reasoned decision-making process. Humans lack this capacity, but robots will be explicitly programmed this way, and thus could be the first truly moral agents. Blay Whitby (2011) wonders, "Is humanity ready or willing to accept machines as moral advisors?" Still others believe that emotions implanted into machines could allow them to become moral agents, but these thinkers are concerned that such a move would mean smart machines would be capable

[3] John-Stewart Gordon (2016) paraphrases this Aristotelian viewpoint as follows: "The fine or the noble and the just require the virtuous person to do or refrain from doing certain things, for example, not to murder (in particular, not to kill one's parents), not to commit adultery, and not to commit theft."

of suffering, thus increasing the cumulative amount of suffering in the world (Scheutz 2012). To crown it all, David J. Calverley (2011) holds that machines could be granted legal rights in the same way that corporations and other non-human entities currently are. The government of South Korea is developing a Robot Ethics Charter which aims to protect humans from abuse by robots—and vice versa (BBC 2007).x

As we see it, machines are indeed increasingly intelligent, but do not currently have (nor will they gain, at least in the foreseeable future) the attributes essential for becoming moral agents. Calls to either implant in these machines a moral philosophy that enables specific moral decisions (the top-down approach) or to let them learn from humans which ethics to follow (the bottom-up approach) presume that machines are or can be made into moral agents. However, machines are not moral agents and there is no indication that they can be made into such agents.

15.2 "Autonomous Machines," A Highly Misleading Term

15.2.1 How Autonomous Are Smart Machines?

In many discussions of the ethical challenges posed by driverless cars, and smart machines generally, they 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 (see for example van Inwagen 2003; Harris 2011). There is no need here to repeat the arguments against this position, and 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 chapter 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 diverted from its original course by a strong gush of wind "decided" to change direction, but this would be merely a misperception, an 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 behavior 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 the inexperience of the person using it, poor eyesight, or some other such 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 such 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, to draw conclusions on its own—such as a more advanced GPS system, which identifies upcoming traffic, or an accident, and reroutes accordingly. Machines equipped with AI are held to be able to act much more autonomously than those not so equipped.

Monica Rozenfield (2016) 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.

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 command full autonomy. For instance, weapon systems that choose their own targets, without human intervention, excluding even the ability to abort the mission. 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 second hand. Military ethicist George Lucas Jr. (2013) 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. Domingos (2015, p. 283) writes:

> They can vary what they do, even come up with surprising plans, but only in service of the goals we set for 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.

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

For the sake of the following discussion, we shall assume that smart machines have a significantly greater capability of rendering their own cognitive choices than old fashioned ones; e.g. deciding 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, all the while not necessarily having an ability to make ethical choices autonomously, as we shall see. Machines are ultimately tools of the human beings who design and manufac-

ture them. If humans fail to include sufficient controls in the construction of machines with AI, then these humans must bear the responsibility of any ill effects—just as if an owner of a poorly trained bulldog cuts his leash, he is responsible for any havoc then wreaked.

15.2.2 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 as a leading principle: physicians and other health care personnel should first and foremost heed the preference of the patient. However, cars and other machines are not emotional beings that experience pain, 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 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 high 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 reeducation—but often, society will also set new limits on them (curfew for teenagers, 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 people are, there appear to be no moral injunctions against implementing extensive constraints on a smart car's behavior. Quite simply, the car is a malfunctioning tool that should be dealt with accordingly.

15.2.3 Partners, Not Free Standing Agents

Another major source of the misconception that seems to underlie much of the discussion is to be found in public discussions of AI, even among academics: 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 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 full 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, 2012).

Those who seek to call attention to the key difference under discussion have used a wide variety of other terms. Some refer to AI minds as "strong AI" (Encyclopaedia Britannica 2016). AI partners have been called Intelligence Augmentation (IA) (Markoff 2015), "intelligence amplification," "cognitive augmentation," or "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.) Many 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[4] 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 pedals 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

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

acts. One could no longer consider only the programmers, manufacturers, and 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 while even the smartest machine is very good at carrying out some functions that humans used to perform (e.g. memorizing), it remains 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 feel fear.

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 any machine. The chapter 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.

15.3 The Main Ethical Implementing Factors: Legal and Personal

How can driverless cars and other such machines follow the ethical preferences of humans if these machines are not provided a capacity to make ethical decisions on their own? In answering this question, one must first consider the two different ways that moral and social values are implemented in the human world; then, how these values might be introduced into the realm of the new 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 regulations; those who do not heed these values are physically prevented from continuing (e.g. their cars are towed if they park illegally), penalized (e.g. issued tickets), or jailed (e.g. drunken drivers). Other values are fostered through informal social controls. Those who violate these values are shamed or chastised away, while those who abide by them are commended and appreciated. Stopping to help stranded motorists is a case in point. Thus, while Good Samaritan acts are required by law in France, in the US they are merely encouraged. Other personal choices include whether one buys environmentally-friendly gasoline or the cheapest available; purchases cars that pollute less than others; stops for hitchhikers; and allows friends to use one's car.

The distinction between the two modes of implementing social and moral values—between legal enforcement and informal social controls—is critical because the *many values* that are implemented through laws enacted by legislatures, inter-

preted by courts, and enforced by the state—*are in principle not subject to individual deliberations and choice!* They are subject to communal 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 though 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 these many important matters, *what cars ought to do is not up to them any more than it is up to their owners*. (True, people may judge the state-imposed rules as unethical and engage in civil disobedience. We take it for granted that, for the time being, we need not concern ourselves with whether cars should or should not be programmed to be able to rebel.) In contrast, there are relatively few important ethical decisions not enshrined in law and are left to individuals to make, at least when one deals with cars, robotic surgeons, and killing machines.

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 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.

A leading AI researcher pointed out that some of these legal measures are not as sharply edged as those listed above, comparing the yield sign to the stop sign. The ambiguity, he argued, leaves room and need for ethical judgments—and hence for cars to be able to make ethical judgments even in cases where there is legal prescription. There is no denying that the law does not define yield signs as precisely as stop signs (which makes programming the car more difficult). Moreover, as Noel Goodall points out, people have a sense of judgment as to when to abide by the spirit of the law rather than its letter, e.g. when to cross a double yellow line when they face an object in their lane and the adjacent one is empty (Goodall 2016). Driverless cars might barrel ahead or else stop suddenly. However, given that computers have been programmed to play chess—which entails taking into account moves by others and responding to these moves—it seems reasonable that cars can be programmed to deal with issues much more complex than those posed by yield signs and other ambiguities in the law. Or, they may be programmed to deal with such situations with less flexibility than humans; say, always wait until other cars are at least five car lengths away (at a given speed) and more—if the speed is higher.

There remain those ethical decisions that are not prescribed by law and which thus must be rendered by an individual or—their car. 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 (see Beshears et al. 2009; Benartzi and Thaler 2013) to organ donations (Johnson and Goldstein 2003, pp. 1338–9) 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 versus 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 own computers) about the acts of a particular individual in order to determine that person's moral preferences. *Essentially, what ethics bots do for moral choices is rather similar to what many AI programs do for ferreting out consumer preferences and targeting advertising to them accordingly, except that in this case, the AI program is used to guide instruments that are owned and operated by the person, in line with their values,* rather than by those of some marketing company or political campaign seeking to advance their own goals. 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. Note that this program does not seek to teach the car an ethical system that will allow it (and other smart machines) to deliberate and then form their own moral conclusions. Rather, it extracts specific ethical preferences from an owner and subsequently applies these preferences to the operations of the owner'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).

A more advanced ethics bot could be used to monitor the behavior of smart machines and rein them in if need be. For instance, we noted that a smart car might originally be instructed not to exceed speed limits, but being a learning machine, stray from these instructions when it imitates other cars that do exceed the speed limit. An ethics bot would monitor the performance of the car, and if the car learned to speed, the bot would reset the car's speed controls, alert its owner, and if necessary, "patch" the car's program. That is, instead of treating the AI world as if it were one unitary field, one should view AI along the same lines as the rest of the world, which is managed in two (or more) layers. Workers have supervisors, teachers have principals, and businesses have auditors as a second layer. This serves to ensure that the performance of the first layer stays within the boundaries of whatever program is required by the mission or undertaking to be carried out, including its ethical standards. (In the human world one recognizes that these supervisors themselves may stray or act unethically, and thus require additional layers of guidance or counterbalancing forces.)

The same structure is now needed for AI: the ever-growing number of smart instruments needs a second line of AI programs to act as supervisors, auditors, accountants—as ethics bots—of the first line AI programs to keep them from straying (See Etzioni and Etzioni 2016a, b).

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 simply takes the moral positions of the human owner and instills them in the machine. One may wonder, 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 truly harmful, and hence remain without regulation or attending legislation.

We have seen so far that implanting 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 laws and regulations. These choices are made for the machines by the community, 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.

15.4 The Outlier Fallacy

A surprisingly great amount of attention has been paid to the applications of Trolley narratives 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 narratives are not without merits. Like other mental experiments, they make for an effective 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 tales 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, making the outcome of each option knowable with 100% accuracy. 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, for example. James O'Connor (2012) adds significantly:

> 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.

Barbara Fried (2012) 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.

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 someone 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.

To reiterate, most of the time, smart machines can be kept in line through legal means. In other situations, they will abide by their owners' ethical preferences (both assisted by second order, supervisory AI programs). Granted, these human guidance modalities will leave "uncovered" the once-in-a-million-miles situations (each unique and often unpredictable). Indeed, one must assume that such events will be even less likely for smart cars because they can respond much more quickly than humans, and thus, for instance, determine if there is enough space between the child and the cars in the adjacent lane to squeeze by. If instead, one made the cars egoisti-

cal (always acting to maximize the safety of the passenger), society is sure to condemn those who make them and drive them, if not ban them altogether. If cars are made to be always altruistic and self-sacrificing, very few would purchase them. However, there will always be some incidents that cannot be foreseen and programmed; this happened, for instance, when a metal 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 programs will need to be updated by humans.

All 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. The community can set legal limits on what these machines do in most cases; their owners and users can provide them with ethical guidance in other situations, employing ethics bots to keep AI equipped machines in line; and collectively, we can leave be the one-in-a-million situation, without neglecting to cover the harm.

Those concerned with these matters might give the Trolley cases a rest and focus instead on ethical issues raised by the proliferation of smart machines that are almost immeasurably more consequential, such as the number of lives saved by the accelerated development of driverless cars of the partner kind. Also, the issues raised by a particular subset of "autonomous" machines—such as killing machines or smart weapons—that should not be allowed to become driven by AI minds but only by AI partners, thereby always keeping a human in the loop who can abort missions and otherwise keep smart machines within bounds.

References

Anderson, S.L. 2011. Philosophical concerns with machine ethics. In *Machine ethics*, ed. Michael Anderson and Susan Leigh Anderson, 162–167. Cambridge: Cambridge University Press.

Anderson, M., and S.L. Anderson, eds. 2011. *Machine ethics*. Cambridge: Cambridge University Press.

Batavia, P.H., D.A. Pomerleau, and C.E. Thorpe. 1996. *Applying advanced learning algorithms to ALVINN*. Pittsburgh: Carnegie Mellon University, The Robotics Institute.

BBC News. 2007, March 7. Robotic age poses ethical dilemma.

Benartzi, S., and R.H. Thaler. 2013. Behavioral economics and the retirement savings crisis. *Science* 339 (6124): 1152–1153.

Beshears, J., J.J. Choi, D. Laibson, and B.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*, 167–195. Chicago: University of Chicago Press.

Bojarski, M. et al. 2016. End to end learning for self-driving cars.

Bonnefon, J.F., A. Shariff, and I. Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352 (6293): 1573–1576.

Calverley, D.J. 2011. Legal rights for machines: Some fundamental concepts. In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 213–227. Cambridge: Cambridge University Press.

Coeckelbergh, M. 2009. Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society* 24 (2): 181–189.

Constant, B. 1797. Des réactions politiques. *Oeuvres complètes* 1: 1774–1799.
DMello, A. 2015. Rise of the humans: Intelligence amplification will make us as smart as the machines. *The Conversation.*
Domingos, P. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.
Encyclopaedia Britannica. 2016. "Artificial Intelligence (AI)."
Etzioni, A., and O. Oren Etzioni. 2016a. AI assisted ethics. *Ethics and Information Technology* 18 (2): 149–156.
———. 2016b. Keeping AI legal. *Vanderbilt Journal of Entertainment & Technology Law* 19 (1): 133–146.
Frankfurt, H.G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66 (23): 829–839.
Fried, B.H. 2012. What does matter? The case for killing the trolley problem (or letting it die). *The Philosophical Quarterly* 62 (248): 505–529.
Gibbs, S. 2015. What's it like to drive with Tesla's Autopilot and how does it work? *The Guardian*.
Goodall, N. 2014. Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board* 2424: 58–65.
Goodall, N.J. 2016. Can you program ethics into a self-driving car? *IEEE Spectrum* 53 (6): 28–58.
Gordon, J.S. 2016. Modern morality and ancient ethics. *Internet Encyclopedia of Philosophy*.
Hargrove, E.C. 1992. *The animal rights/environmental ethics debate: The environmental perspective*. Albany: SUNY Press.
Harris, S. 2011. *The moral landscape: How science can determine human values*. New York: Simon and Schuster.
Harris, M. 2015. New pedestrian detector from Google could make self-driving cars cheaper. *IEEE Spectrum.*
Hew, P.C. 2014. Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology* 16 (3): 197–206.
Himma, K.E. 2009. Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology* 11 (1): 19–29.
Hsu, J. 2016. Deep learning makes driverless cars better at spotting pedestrians. *IEEE Spectrum*.
Johnson, D.G. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology* 8 (4): 195–204.
Johnson, E.J., and D. Goldstein. 2003. Do defaults save lives? *Science* 302 (5649): 1338–1339.
Joy, B. 2000. Why the future doesn't need us. *WIRED* 8: 238.
Levin, S., and N. Woolf. 2016. Tesla driver killed while using autopilot was watching Harry Potter, witness says. *The Guardian.*
Lohr, S. 2015. Homes try to reach smart switch. *New York Times.*
Luban, D. 2005. Liberalism, torture, and the ticking bomb. *Virginia Law Review* 91: 1425–1461.
Lucas, G.R., Jr. 2013. Engineering, ethics and industry: The moral challenges of lethal autonomy. In *Killing by remote control: The ethics of an unmanned military*, ed. B.J. Strawser, 211–228. New York: Oxford University Press.
Luciano, F., and J.W. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14 (3): 349–379.
Markoff, J. 2015. *Machines of loving grace: The quest for common ground between humans and robots*. New York: ECCO.
McDermott, D. 2011. What matters to a machine. In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 88–114. Cambridge: Cambridge University Press. 88–114.
Metz, C. 2016. Self-driving cars will teach themselves to save lives—But also take them. *The Atlantic*.
Mill, J.S. 2008. *On liberty and other essays* (originally published in 1859). Oxford: Oxford University Press.
Millar, J. 2014. You should have a say in your robot car's code of ethics. *WIRED*.

Moor, J. 2011. The nature, importance and difficulty of machine ethics. In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 13–20. Cambridge: Cambridge University Press.

Nadeau, J.E. 2006. Only androids can be ethical. In *Thinking about android epistemology*, ed. K. Ford and C. Glymour, 241–248. Cambridge, MA: MIT Press.

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.

O'Connor, J. 2012. The trolley method of moral philosophy. *Essays in Philosophy* 13 (1): 242–255.

Rozenfield, M. 2016. The next step for artificial intelligence is machines that get smarter on their own. *The Institute*.

Scheutz, M. 2012. The affect dilemma for artificial agents: Should we develop affective artificial agents? *IEEE Transactions on Affective Computing* 3 (4): 424–433.

Shah, R.C., and C. Sandvig. 2008. Software defaults as de facto regulation the case of the wireless Internet. *Information, Community & Society* 11 (1): 25–46.

Sharkey, N., and A. Sharkey. 2010. The crying shame of robot nannies: An ethical appraisal. *Interaction Studies* 11 (2): 161–190.

Sharkey, A., and N. Sharkey. 2012. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology* 14 (1): 27–40.

Spice, B. 2016. *Carnegie Mellon transparency reports make AI decision-making accountable*. Carnegie Mellon Computer University School of Computer Science.

van Inwagen, P. 2003. Fischer on moral responsibility. *The Philosophical Quarterly* 47 (188): 373–381.

Wallach, W., and C. Allen. 2009. *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Whitby, B. 2011. On computable morality: An examination of machines. In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 138–150. Cambridge: Cambridge University Press.