AI assisted ethics

Amitai Etzioni & Oren Etzioni

Ethics and Information Technology

ISSN 1388-1957
Volume 18
Number 2

Ethics Inf Technol (2016) 18:149-156
DOI 10.1007/s10676-016-9400-6

Springer

Ethics Inf Technol (2016) 18:149–156
DOI 10.1007/s10676-016-9400-6

ORIGINAL PAPER

AI assisted ethics

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

Published online: 5 May 2016

Abstract The growing number of 'smart' instruments, those equipped with AI, has raised concerns because these instruments make autonomous decisions; that is, they act beyond the guidelines provided them by programmers. Hence, the question the makers and users of smart instrument (e.g., driver-less cars) face is how to ensure that these instruments will not engage in unethical conduct (not to be conflated with illegal conduct). The article suggests that to proceed we need a new kind of AI program—oversight programs—that will monitor, audit, and hold operational AI programs accountable.

Keywords Ethics bot · Communiterianism · Second-layer AI · Driverless cars

Introduction

The question of which values should be introduced into the guidance systems of driverless cars has implications well beyond the ethical directions to be granted to these new vehicles. Namely, such guidance is needed for a great variety of robots, machines, and instruments (instruments, from here on) that are already equipped with artificial intelligence (AI)—and many more in the near future (The Economist 2015). These instruments are often referred to as "smart." As Ed Lazowska of the University of Washington put it, "During the next decade we're going to see smarts put into everything. Smart homes, smart cars, smart health, smart robots, smart science, smart crowds and smart computer–human interactions" (Markoff 2013). According to Francesca Rossi, a computer scientist at the University of Padova, "Until now, the emphasis has been on making machines faster and more precise—better able to reach a specific goal set by humans. Today, the aim should be to design intelligent machines capable of making their own good decisions according to a human-aligned value system" (Rossi 2015). Gary Marcus of New York University holds that in the near future a moment will arrive that will herald an "era in which it will no longer be optional for machines to have ethical systems" (Marcus 2012).

One should note, a note essential for all that follows, that these smart instruments are able not only to collect and process information in seconds much more efficiently than human beings can do in decades or even in centuries—but also to form decisions on their own. That is, AI provides these instruments with a considerable measure of autonomy in the sense that they often will not inquire of their human users how to proceed and instead will render numerous decisions on their own (Mayer-Schönberger and Cukier 2014: 16–17). Stuart Russell discusses the development of algorithms that closely "approximate" autonomous human behavior and values (Wolchover 2015). Autonomy in computer science thus refers to the ability of a computer to follow a complex algorithm in response to environmental inputs, independently of real-time human input. That is, autonomous robots are "robots that can figure things out for themselves" (2015). For instance, self-driven cars decide when to speed up or slow down, when to hit the brake, how much distance to keep from other cars and so on.

✉ Amitai Etzioni
etzioni@gwu.edu

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

[2] The Allen Institute for Artificial Intelligence, Seattle, USA

It follows that if these smart instruments are not to act like amoral machines, their AI guidance programs will need to include substantial moral components.[1] To put it differently, given that decisions tend to have a moral dimension (Etzioni 1988)—the programs that guide all these smart instruments need moral programming (Rossi 2015; Tegmark 2015). For instance, several scholars have asked under what conditions driverless cars would be instructed to swerve into a parallel lane to avoid hitting a kitten—even if such a move could cause several human fatalities (Marcus 2012; Bonnefon et al. 2015). The same question has numerous permutations, such as whether a car on a busy road should swerve to avoid hitting a child or adult if doing so would risk causing a pileup that could kill several people, or whether it should swerve to avoid hitting a non-living but solid obstacle to protect its own occupants even if such a move will lead to hitting a car in another lane. Moreover, driverless cars will need to be instructed whether they should slow down in order to stop when they see a hitchhiker or a car accident down the road, how to react to the road rage of a driver of an old fashioned car, whether to join a long queue or try to cut in, and many other such value-laden questions.

These questions may remind readers of the moral dilemma involving a trolley coming down a track and a person at a switch who must choose whether to let the trolley follow its course and kill five people or to redirect it to another track and kill just one. (This rather popular mental experiment has been used in several permutations.)[2] However, the subject at hand is rather different because the trolley decisions are made by a person; when AI is used many decisions are made often by the instruments themselves. These programs are correctly referred to as "black boxes" and as lacking accountability and even "traceability" (Mayer-Schönberger and Cukier 2014: 141, 178). Hence, they need to be given a priori and continuous moral guidance if they are to heed the values of their users and of the community.

The article next turns to examining several suggestions that have been made about the way moral guidance is to be provided to smart instruments, those equipped with AI, and then adds a distinct approach.

Social moral values ensconced in law

One major answer to the question of which values should be embodied in AI guidance systems is that the values shared by a particular community should be used. For many issues, this community would be the nation augmented by local communities (in the United States, these would be states and municipalities). Obvious examples are values such as thou shalt not kill, steal, rape, harm others, or harm the environment.

Thus, the guidance systems of driverless cars will be expected to ensure that these cars observe various speed limits, keep a safe distance from other cars, and so on as the drivers of old-fashioned cars are required to do. In short, one part—the easy part—of the answer to the question of which values are to be implemented in the guidance systems of smart instruments is: the values ensconced in the law of the community or communities in which the smart instruments are employed.

Several questions, though, remain even about the values embodied in law. First, who should be charged if instruments violate the law? The owner, the user, the designer, the manufacturer—or the computer that in effect operates the instruments and renders many decisions on its own (Kaplan 2015)?[3] The law clearly treats cases in which there was intent to cause harm much more harshly than cases in which there was no such intent. Compare the ways the law treats murder and involuntary manslaughter. (The question of intent also figures in assigning liability.) But how is one to determine whose intent (if any) was the cause when one cannot trace the process by which the decisions were made?

[1] Stuart Russell of University of California Berkeley stated, "You would want [a robot that does everyday activities] preloaded with a pretty good set of values" (Goldhill 2015).

[2] Philosophers Philippa Foot and Judith Jarvis Thomson described a situation called "the Trolley Problem," which raises the question whether a runaway train is about to run over a group of five people on the tracks, but their deaths could be averted by flipping a switch that would divert the train onto another track, where it would kill one person (Lin 2013).

Joshua D. Greene describes a number of ethical dilemmas that generally fit into the category of the "trolley problem." These include "switch" cases, in which throwing a switch will turn the trolley away from some number of people toward a single person, or "footbridge" cases in which one must push a person into the path of the trolley to save others' lives. He also discusses similar famous ethical questions studied by researchers, such as the question whether it is immoral to allow a child to drown in a shallow pond to avoid muddying one's clothes, whether one is morally obligated to donate money to save others' lives, and more (Greene 2014).

Jean-Francois Bonnefon, Azim Shariff, and Iyad Rahwan apply this question to the issue of driverless cars; as driverless cars ("autonomous vehicles," or AVs) and other forms of use of artificial intelligence become more widespread. They examined whether individuals would be comfortable with AVs programmed to be utilitarian and found that the answer was generally yes (Bonnefon et al. 2015).

[3] One scholar at the National Science Foundation points out that technology currently outstrips knowledge of how to assign liability for robots' ethical and legal failures (The Economist 2014).

Professor Patrick Lin points out that algorithms cannot make "an instinctive but nonetheless bad split-second decision" the way humans can, and thus the threshold for liability may be higher (Lin 2013).

Second, what level of law enforcement does society seek, given that smart instruments make a high level of surveillance very easy to achieve? (For instance, one could determine from a central location whether truck drivers are driving too long without taking a break, the speed of any car on the road, and the location of all who use cell phones and much more.) Third, what guidance is to be given to instruments when the law and ethics diverge? (For instance, should driverless cars be programmed to refuse to violate the law even when the driver takes over, or to allow speeding in an emergency?) Wrestling with these questions is left for a separate discussion in order to focus here on the social and moral values not ensconced in law (Etzioni and Etzioni 2016).

A communitarian approach

Numerous social and moral values are not ensconced in law. These include the extent and scope of one's commitments to one's children, spouses, friends, neighbors, the various communities to which one belongs, the nation, and even the international community (Wrong 1995). These values include taking risks for others (such as fighting overseas and donating organs), volunteering, giving to charity, resolving differences with others civilly, and many others. Indeed, many values ensconced in law are paralleled by considerable additional moral commitments above and beyond those required by law. For instance, most of what the moral culture of communities expects parents to do for their children greatly exceeds what the law commands. A communitarian position holds that it is essential to include these values in the guidance systems of smart instruments because these values make for a good, civil society well beyond a stable and even liberal state.

How should one determine which communal values to incorporate into the AI guidance systems of smart instruments? Those communal values ensconced in law can be identified relatively readily—they are values that legislatures and courts draw upon when they enact laws and interpret them. But how is one to determine which additional social moral values the community seeks to foster?

Some suggest that these additional values should be those shared by the community (Walzer 1984). This position runs into several difficulties. First, people belong to different communities and to the encompassing society, which subscribes to different values. For instance, famous attempts to use community standards to determine obscenity failed because of disagreement about what behavior qualifies (Jacobellis v. Ohio 1964). Even when there is considerable consensus about values at a high level of abstraction, there is often much less consensus about what specifically these values require. Thus many people agree that the environment ought to be protected, but disagree about which mileage standards cars should have to abide in by 2030, what level of emissions they should be allowed to produce, whether one should idle at stop signs, and whether they should drive slower than the law requires. This poses great difficulties for programmers.

A related question is how to determine what the values are of whatever community the instruments are to heed. Several scholars have suggested using focus groups or public opinion polls to determine what the relevant values are.[4] One notes, however, that the results of public opinion polls vary significantly depending on who is surveyed, question wordings, the sequence in which questions are asked, the context in which questions are asked (e.g., at home versus at work), and the attributes of those who ask the questions (e.g., are they the same race as the person queried). Even when the same question is asked of the same people by the same people twice, rather different answers follow (Institute for Statistics Education). Hence such polls cannot be used as a reliable base.

The suggestion that the time has come to engage in what might be called "teaching machine ethics," that is, teaching instruments to render moral decisions on their own (Lin 2013), runs into different difficulties. Driverless cars and other such instruments are unable to form moral guidelines out of thin air. They will need, at least to begin with, first principles and some guiding philosophy, say a utilitarian or a deontological one. After all, even humans do not start with a tabula rasa. They gain ethical foundations from those that raise them and from their communities and then modify or replace these foundations over time. Which principles and methodologies should be given to smart instruments? Is there a reason they should all be given the same foundations? Could cars come with ethical options—some equipped with utilitarian principles, or deontological ones, or Buddhist philosophies? (One may wonder whether buyers would understand the implications of their choice of car unless they took some philosophy classes.) Some kind of polling approach may be unavoidable when deciding whether or not to use instruments that are used by communities rather than individuals. For instance, such an approach may be necessary when determining whether to post in public spaces cameras equipped with AI systems that scan the footage to identify people who act in an uncivil manner. However, there seem to be serious

[4] Slobogin and Schumacher (1993: 757) recommend that the Supreme Court draw on public opinion polls to determine that about societal expectations of privacy. Similar suggestions were made by Fradella, Morrow, Fischer, and Ireland. They conducted a survey of 589 individuals (Fradella et al. 2010–2011: 293–94).

Bonnefon et al. (2015) applied this idea to finding which values ought to guide self driving cars.

difficulties associated with relying on a communitarian approach to determining which values the instruments should embody for most instruments that are owned and operated by individuals. Is there no way to more closely align the values that instruments are expected to heed and the ethics of their intended users?

A libertarian approach

Libertarians and some liberals hold that each person should define the good and the values they are to heed, and that the state should remain neutral (Boaz 1999). It would be compatible with this position if smart instruments came with a rich menu, which would allow each individual to choose which options are in line with their values, as well as an opened ended category which would allow them to include attention to moral preferences not included in the menu.

The difficulties this libertarian approach raises are illustrated by the development of privacy options. Many websites initially had merely a statement of the privacy policy they follow, which put the users in a "take it or leave it" position—assuming they understood the legal statements. Next, an increasing number of websites offered users a small menu of choices regarding the level of privacy they preferred. Facebook, for instance, offers five main privacy settings. Even at this level, people complained about the complexities of these settings. A Google representative recently stated that Google would provide up to a hundred such options (Fleischer 2015). Strong evidence from psychological studies and experience suggests that most users will find the requirements to make that many choices on their own overwhelming (Kahneman 2011).

These difficulties are much more challenging if one takes into account that individuals would need to personally provide individual moral guidance to all the growing number of smart instruments one uses as the world is moving into the "internet of things." In short, this approach seems highly impractical.

AI assisted ethics (ethics bots)

A paradigmatic agenda

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 what the acts of a particular individual that reveal that person's moral preferences are*. Basically 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. 5 Only in this case, the bots are used to guide instruments that are owned and operated by the person, in line with their values, rather than by some marketing company or political campaign seeking to advance their goals. For instance, an ethics bot may conclude that a person places high value on environmental protection if the ethics bot finds that the 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 purchase only environmentally friendly fuels, to turn on the air conditioning only if the temperature is high, and to idle the engine at stops.

Much of what follows about other attributes of ethics bots is *paradigmatic*, in the sense that the article outlines the qualities of these as-yet to be developed ethics bots. Some readers may well consider the following pages somewhat visionary in terms of what they assume AI will be able to accomplish in the not-too-distant future. However, given the inability to implement communitarian and libertarian approaches, it seems better to employ even weak ethics bots, at least initially, than to continue to leave driverless cars and the large and growing number of other smart instruments without ethical guidance. Moreover, we shall see that some, albeit rather simple, ethics bots have already been constructed, road tested, and used by a considerable number of people.

To illustrate: nest built a smart thermostat. It first "observed" the behavior of the people in various households for merely a week and drew conclusions about their preferences. It then used a motion-detecting sensor to determine whether anyone was at home. When the house was empty, the smart thermostat entered a high energy-saving mode; when people were at home, the thermostat adjusted 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 moral values. This thermostat enables people with differing values to have the temperature settings they prefer. The residents of the home do not need to reset the thermostat every day when coming and going. This simple ethics bot also reduces the total energy footprint of the community (Lohr 2015: 147).

The ethics bot so far depicted is the most basic version. Additional features and more sophisticated ethics bots that might be developed, are outlined below. They all share, though, two major features. First, they enable the people who *employ AI to guide AI*. This is, instead of treating the AI world as if it were one unitary field AI should be

restructured along the same lines as the rest of the world is. That is, some AI programs should serve as the first line "worker bees" that provide directions to an ever growing number of instruments—from robot soldiers to Barbie dolls, from voter mobilization drives to refrigerators. Second line AI programs will act as supervisors, auditors, accountants, and as ethics bots of the first line AI programs.

Second, because people bring their same basic values to different pursuits, once an ethics bot is able to carry out an analysis of the moral preferences of a person, the same findings will help guide a variety of smart instruments the person uses. Thus, a person does not need one ethics bot for shopping, another for driving, and still another for volunteering. For example, if an ethics bot determines that a given person's moral preferences are to maximize their self-interest, that bot would instruct the person's instruments to shop at places they find the lowest costs and best quality but disregard whether the sellers have been charged with employing workers at unsafe locations overseas; paying less then minimum wages; and polluting. It would also instruct the person's financial AI system to make donations to charity only if those donations generate enough deductions to make up for the "loss" or if they engender for the donor a great deal of goodwill. And so on.

Of course, the ethical preferences of most people are more complicated than the simple examples here used the purposes of exposition. Hence, in the longer run, ethics bots would need to be similarly internally diverse. Moreover, ethics bots of the future should be able to self-update at regular intervals in order to take changes to people's moral preferences into account. Last but not least, ethics bots should have an override feature, discussed below.

Basically 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.[5] Only in this case, the bots are used to guide instruments that are owned and operated by the person, in line with their values, rather than by some marketing company or political campaign seeking to advance their goals.

Ethics bots, once developed, should be able to provide a superior interface between a person and smart instruments compared to unmediated interaction. There are several reasons for this. First, if most people will be required to render a large number of ethical choices, they will quickly give up because a sort of psychological fatigue sets in akin to the one felt by people who are trying to consider their best chess move and after a while just give up and move. (By contrast, AI is very patient, which is one reason it now beats even the world chess masters.)

Second, people accommodate their inability to make a great number of choices by making lexicographic choices.[6] That is, they ignore the information about most facets of the object of their choice, and focus on a few that they hold to be most important. Thus, when they buy a car, they may examine its relative price, miles per gallon, and color, or some other such mix of features—but ignore scores of other attributes. The same holds true of moral choices. People, when making a major donation, may take into account the goals the given charity serves, whether it services people in their own community or overseas, and whether it has a reputation as an honest agency, or some other such mix. They will ignore, in the process, many other features of the given charity such as its long-term record, recent changes in leadership, its ratio of expenses to payouts, and so on. In contrast, ethics bots, given their computing power, have no such limitations. They hence will help people to ensure that their choices about how and where to donate, shop, vote and more much more closely reflect all their moral preferences than a few selected ones.

Third, ethics bots are likely to compare favorably to other means—such as interviews, self-administered forms, and mental exercises such as those used in lifeboat ethics[7]—that seek to ferret out a person's moral preferences. This is largely the case because these subjective means draw their conclusions mainly on the basis of *expressed attitudes*, while ethics bots mainly note the moral choices revealed in *actual behavior*. For instance, one may say that one attends church regularly, but an ethics bot would note that the person played golf often at the time religious services are carried out and parked at the place of worship only a few times a year. This attribute is of special importance because attitudes, a great deal of data shows,

[5] For example, Nielsen has developed a marketing system for targeting very specific demographics with financial and investment products based on age, affluence, the presence of children in the home, and certain purchasing habits. These include such specific target consumer groups as "Y2-54: City Strivers" and "F4-56: Economizers" (Nielsen 2015); Ted Cruz' campaign in Iowa relied on psychological profiles to determine the best ways to canvass individual voters in the state (Hamburger 2015).

[6] Lexicographic preferences are those in which "respondents have a ranking of the attributes [consider important], but their choice of an alternative is based solely on the level of their most important attribute(s)" (Campbell et al. 2006).

[7] "Lifeboat ethics" refers to an ethical dilemma outlined by Garrett Hardin in 1974, which describes a situation in which a lifeboat nearly full to capacity must consider whether or not to bring aboard ten additional passengers out of 100 people in the water. The purpose of lifeboat ethics-style philosophical discussions is not to tell anyone what is ethically correct in any given situation, but rather to help individuals to clarify their own values.
This is one component of a larger school of ethics, "moral reasoning." Moral reasoning encourages "individual or collective practical reasoning about what, morally, one ought to do" (Richardson 2014).

correlate poorly with behavior.[8] Also, people have difficulties in articulating their preferences.

The introduction of ethics bots would raise serious privacy concerns. Many people may well seek to encrypt them and call for laws that would treat "reading" another person's ethics bot without written prior permission as akin to reading their medical record or other sensitive information.[9] One should note, though, that ethics bots would not be mandated, and hence would be used only by those who see their merits and benefits as exceeding the bots' privacy risks.

Moreover, those who adopt ethics bots are sure to note that most of the information bots use to ascertain their preferences is already publicly available. Several corporations maintain very detailed dossiers on most people and sell these dossiers to all comers. For instance, Axciom maintains dossiers on most Americans. These dossiers include "age, behavior, buying activity, financial, household, interest, real property, life events," and more, up to 1500 items per person. SeisInt dossiers include individuals' "criminal histories, photographs, property ownership, SSNs, addresses, bankruptcies, family members, and credit information."[10] These dossiers can even include sensitive medical information. As Eli Pariser reports, "Search for a word like 'depression' on Dictionary.com, and the site installs up to 223 tracking cookies and beacons on your computer so that other websites can target you with antidepressants" (Pariser 2011). It seems that if others are free to use a very great deal of personal information to promote products and politicians and to seek to shape people's preferences, there is little to be lost and much to be gained if that person uses the same information to ensure that the instruments *they* employ will comport with *their* values.

Individuals can at any point override the guidance an ethics bot provides to their instruments. Thus, when a smart thermostat programmed by Nest did not follow individuals' preferences, but rather set the thermostat within two degrees of their preferences at a setting more favorable to the environment, many people rejected this setting (Lohr 2015). In short, ethics bots (or AI assisted ethics) are a marriage between libertarianism (because the particular person provides the moral preferences—the definition of the good) and AI programs (which provide moral guidance to smart instruments).

Individuals can use their ethics bots for self-assessment

Adolescents, people in psychotherapy, and many others often engage in self-examination, including asking themselves whether they are good people, whether they do enough to serve others, and more. Studies of such self-assessment suggest that people often greatly over- or under-evaluate themselves (Dunning et al. 2003). In the future, these individuals will be able to draw on their ethics bots to provide them a more objective—and candid—evaluations of themselves. These evaluations may well include how they compare to others in their community and whether they improve or not over time. Here ethics bots serve as tools of moral self-improvement.

Ethics bots can be programmed in ways that allow their human users to modify the program the bots chose for them, based on their preferences. They can augment, distract, and override—all moves that are much less taxing than forming the self's ethics profile de novo. For instance, a person may note that his ethics bot reveals that over the last 10 years he donated rather little to various charities then he thought he did—and instruct his financial app to increase these allocations.

Ethics bots can also help people implement pre-commitment strategies. The term refers to taking steps before a situation in which one expected to be tempted to act unethically—to fend off the temptation. Odysseus employed this strategy by instructing his sailors to tie him to the mast of the ship and to plug their ears before they entered the sea of Sirens so that they would not be tempted by the Sirens' calls (Homer 1978). Thus, an ethics bot could be set to instruct a driverless car not to yield to attempts to override the car's system to engage in road rage, or to mute the car's horn if the person blows it long, hard, and often.

Pragmatic and operational considerations

Some critics may well argue that the agenda charted so far is well beyond what AI can achieve. They may point to programs that try to divine consumer preferences much simpler than their ethical preferences and did not fare well, such as, programs that recommend books and movies. In general, arguments about what AI can and cannot accomplish swing between overblown hype and overblown despair. Some point to AI programs that play winning chess, Jeopardy, and Go, while others bemoan the difficulties computers have in accomplishing tasks that are simple to humans, such as reading graphic designs.

In response, one must reiterate that humans *cannot* on their own provide more than very elementary ethical

[8] Across many different situations, it is well-established that "attitudes are poor predictors of behavior" (Ajzen and Fishbein 2005). See also: Azjen et al. 2004.

[9] For an in-depth discussion of the different treatment afforded to less and more sensitive information, see: [redacted].

[10] *See id.*

guidance to smart instruments without AI assistance. Hence, even weak ethics bots seem preferable to none, unless one can come up with a still different way to provide moral guidance to smart instruments. The best test of whether ethics bots can be created is to invest more in trying to build some. There seem to no obvious, a priori, logical reasons to hold that such bots cannot be constructed—and some, albeit very simple ones, have been developed.

Determining the moral preferences of a person in some areas may well be less daunting than in others. For instance, determining what a person considers fair might be indeed very difficult. By contrast, people's privacy preferences seem easier to grasp. Most people are either privacy fundamentalists (as their use of multiple personal email addresses, frequently changing their passwords, and so on reveals); privacy pragmatists (who will share personal information if the price is right); or privacy unconcerned (Westin 2003). In short, ethics bots are badly needed. Whether more and more accomplished ones can be constructed in the near future remains to be seen. Meanwhile, driverless cars are roaming the streets and so far have not been granted any moral guidance. Even if we cannot be able to construct a high-fidelity model of people's ethical preferences, we will be able to approximate and the approximation will improve as the technology gets better.

In conclusion

The incorporation of AI into more and more instruments makes them much smarter—more efficient, and more effective. In the process, these instruments are acquiring a measure of autonomy in the sense that they render many decisions on their own, well beyond the guidelines that the programmers introduced and sometimes even counter to these guidelines. There is hence growing concern about how society and millions of individuals can rest assured that instruments they use, which are equipped with AI, will not render unethical decisions (Dellinger 2015).

The answer to these challenges cannot be found in bare human controls, because human beings cannot determine on their own whether unethical (and even illegal) acts carried out by smart instruments were the results of the human programming—or AI processes "under the hood." Hence, a major conclusion of this article is that *to ensure proper conduct by AI instruments—people will need to employ other AI systems. Implementing ethical preferences, when dealing with smart instruments, will need to be AI assisted.*

These second order AI programs would have to vary a great deal from one another. For instance, those second order programs that would ensure that instruments do not violate laws, the dictates of which are relatively clear, are likely to differ greatly from those second order programs that would direct instruments to heed moral values, which are often fuzzy. Whether a particular instrument is used by individuals or by a community is also be a factor. This article focuses on those AI programs dealing with compliance with social and moral values for instruments used by a person, such as driverless cars.

The article finds that relying on guidance in these matters on values shared by this or that community raises many difficulties. The same holds if one seeks individuals to directly instruct their smart instruments on their own. A preferred method, here outlined, is to develop AI programs to be used to determine the moral preferences of people, ethics bots, and for these ethics bots to guide the smart instruments.

References

Ajzen, I., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Azjen, I., Brown, T. C., & Carvajal, F. (2004). Explaining the discrepancy between intentions and actions: The case of hypothetical bias in contingent valuation. *Personality and Social Psychology Bulletin, 30*(9), 1108–1121.

Boaz, D. (1999). Key concepts of libertarianism. Cato Institute. January 1. http://www.cato.org/publications/commentary/key-concepts-libertarianism.

Bonnefon, J., Shariff, A., & Rahwan, I. (2015). Autonomous vehicles need experimental ethics: Are we ready for utilitarian cars? *Computers and Society*. October 12. http://arxiv.org/abs/1510.03346.

Campbell, D., Hutchinson, W. G., & Scarpa, R. (2006). Lexicographic preferences in discrete choice experiments: Consequences on individual-specific willingness to pay estimates. Working Paper, Fondazione Eni Enrico Mattei. http://agecon search.umn.edu/bitstream/12224/1/wp060128.pdf.

Dellinger, A. J. (2015) Tim Wu says Google is degrading the Web to favor its own products. *The Daily Dot*. June 29. http://www.dailydot.com/technology/google-search-tim-wu-yelp/.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*(3), 83–87.

Etzioni, A. (1988) *The moral dimension*. New York: The Free Press.

Etzioni, A., & Etzioni, O. (2016). Keeping AI Legal. *Vanderbilt Journal of Entertainment & Technology Law* (Forthcoming).

Fleischer, P. (2015). Privacy and future challenges. Speech, Amsterdam Privacy Conference. Amsterdam, The Netherlands, October 23–26.

Fradella, H. F, et al. (2010–2011). Quantifying Katz: Empirically measuring 'Reasonable Expectations of Privacy' in the fourth amendment context. *American Journal of Criminal Law 38*, 289–373.

Goldhill, O. (2015). Human values should be programmed into robots, argues a computer scientist. *Quartz*. October 31. http://qz.com/538260/human-values-should-be-programmed-into-robots-argues-a-computer-scientist/.

Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics, 124*(4), 695–726.

Hamburger, T. (2015). Cruz campaign credits psychological data and analytics for its rising success. The Washington Post. December 13, https://www.washingtonpost.com/politics/cruz-campaign-credits-psychological-data-and-analytics-for-its-rising-success/2015/12/13/4cb0baf8-9dc5-11e5-bce4-708fe33e3288_story.html.

Hardin, G. (1974). *Lifeboat ethics: The case against helping the poor*. Psychology Today.

Homer. (1978) *Odyssey* (J. H. Finley, Jr. Trans.). Boston: Harvard University Press.

Institute for Statistics Education. Glossary of statistical terms test–retest reliability. http://www.statistics.com.

Jacobellis v. Ohio. (1964). 378 U.S. 184.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Firrar, Straus, and Giroux.

Kaplan, J. (2015). Who put the robot in charge? *Medium*, September 22. https://medium.com/the-wtf-economy/who-put-the-robot-in-charge-408a47335176#.mb8mqqs9p.

Lin, P. (2013). The ethics of autonomous cars. *The Atlantic*. October 8. http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lohr, S. (2015). *Data-ism: The revolution transforming decision making, consumer behavior, and almost everything else*. London: OneWorld Publications.

Marcus, G. (2012). *Moral machines*. New York: The New Yorker.

Markoff, J. (2013). The rapid advance of artificial intelligence. The New York Times. October 14. http://www.nytimes.com/2013/10/15/technology/the-rapid-advance-of-artificial-intelligence.html?pagewanted=all&_r=0.

Mayer-Schönberger, V., & Cukier, K. (2014). *Big data*. New York: Houghton Mifflin Harcourt.

Nielsen. (2015). Nielsen P$YCLE Lifestage Groups. https://segmentationsolutions.nielsen.com/mybestsegments/Default.jsp?ID=8010&pageName=Learn%2BMore&menuOption=learnmore. Accessed 17 Dec.

Pariser, E. (2011). What the Internet knows about you. CNN. May 22. http://articles.cnn.com/2011-05-22/opinion/pariser.filter.bubble.

Richardson, H. S. (2014). Moral reasoning. *The Stanford Encyclopedia of Philosophy* (Winter Edition), Ed. Edward N. Zalta. http://plato.stanford.edu/entries/reasoning-moral/.

Rossi, F. (2015). How do you teach a machine to be moral? The Washington Post. November 5. https://www.washingtonpost.com/news/in-theory/wp/2015/11/05/how-do-you-teach-a-machine-to-be-moral/.

Science Daily. (2015). New algorithm lets autonomous robots divvy up assembly tasks on the fly. May 27. http://www.sciencedaily.com/releases/2015/05/150527142100.htm.

Slobogin, C., & Schumacher, J. E. (1993). Reasonable expectations of privacy and autonomy in fourth amendment cases: An empirical look at understandings recognized and permitted by society. *Duke Law Journal, 42*, 727–775.

Tegmark, M, et al. (2015). An open letter: Research priorities for robust and beneficial artificial intelligence. *Future of Life Institute*. http://futureoflife.org/ai-open-letter/.

The Economist. (2014). That thou art mindful of him. March 29.

The Economist. (2015). Rise of the machines. http://www.economist.com/news/briefing/21650526-artificial-intelligence-scares-peopleexcessively-so-rise-machines.

Walzer, M. (1984). *Spheres of Justice: A defense of pluralism and Equality*. New York: Basic books.

Westin, A. (2003) Social and political dimensions of privacy. *Journal of Social Issues 59*(2). http://onlinelibrary.wiley.com/doi/10.1111/1540-4560.00072/epdf.

Wolchover, N. (2015). Concerns of an artificial intelligence pioneer. Quanta. April 21. https://www.quantamagazine.org/20150421-concerns-of-an-artificial-intelligence-pioneer/.

Wrong, D. (1995). *The problem of order: What unites and divides society*. Boston: Harvard University Press.