AI and Ethics (2021) 1:27–31
https://doi.org/10.1007/s43681-020-00003-6

OPINION PAPER

AI and ethics

Susan Leigh Anderson[1] · **Michael Anderson**[2]

Received: 10 August 2020 / Accepted: 17 August 2020 / Published online: 6 October 2020

Abstract
Since it is of critical importance that autonomous systems, whether software or hardware, that interact with human beings (and perhaps other sentient beings as well) behave in an ethical manner, we consider six possible approaches to effecting this. We argue that the first five approaches are unsatisfactory and defend the last approach, the approach we have taken. It involves discovering ethically relevant features and corresponding prima facie duties present in the various possible actions such a system could take in particular domains and discovering decision principles for when there is a conflict between those duties. We, further, maintain that there are a number of additional benefits to taking this approach that involve becoming clearer about human ethics, in addition to the ethics to which autonomous systems should adhere, and the chance that it might well lead to providing inspiration for humans to behave more ethically.

Keywords AI · Ethics · Machine ethics · Autonomous systems

1 Introduction

There are many necessary activities that we would like to be able to turn over entirely to autonomously functioning machines, because the jobs that need to be done are either too dangerous or unpleasant for humans to perform, or there is a shortage of humans to perform the jobs, or machines could do a better job performing the tasks than humans. We must ensure, however, that they carry out their tasks in an ethical manner.

For many, ethical issues are thought to only arise in "life or death" situations. We believe that this is incorrect. Whenever the actions of an autonomous system, software or hardware, that interacts with humans (and perhaps other sentient beings as well) could adversely or positively affect them, it is a matter of ethical concern. Since this is the case with each action it takes (even, for example, when an elder-care robot decides to recharge its batteries, because it is not doing something else at that moment that might be ethically preferable), all of its actions should be ethically evaluated.

Ethics is concerned with determining which action or policy would be the best one, given a particular set of circumstances, not just with preventing an undesirable outcome. Therefore, using the primary rule in biomedical ethics ("first, do no harm"), which some have argued for, is not ideal. Consider the example of self-driving cars. Since there are bound to be some accidents, causing harm does that mean that they should not be developed and put into practice? We ought to compare the number of deaths and injuries there are now with human drivers with what would likely happen with only self-driving cars that do not have drunken, texting, and otherwise distracted drivers in control of vehicles to see whether there are likely to be fewer deaths and injuries with self-driving cars.

Since we need to ensure that such systems are used ethically or behave in an ethical manner, let us consider various approaches to effecting this. There would seem to be six general approaches to doing so:

1. We could "hard-code" them so as to prevent them from allowing/performing certain actions that we consider to be unethical.
2. We could put the burden on the user of ensuring that they will only be used ethically, simply providing warnings.

✉ Susan Leigh Anderson
Susan.Anderson@UConn.edu

Michael Anderson
Anderson@Hartford.edu

[1] Department of Philosophy, University of Connecticut, Storrs, CT, USA

[2] Department of Computer Science, University of Hartford, Hartford, CT, USA

3. We could learn from earlier decision-makers' judgements or current polls of what the general populace thinks is ethically acceptable behavior and have this guide such systems.
4. We could use an existing ethical principle or theory to guide the behavior of such systems.
5. We could impose a hierarchy of ethical principles on programs or machines to guide their behavior.
6. We could attempt to learn what is ethically acceptable from those with expertise in ethics, deriving from their input not only ethical principles appropriate for such systems that function in particular domains, but also a way to represent the building blocks of ethics.

Let us consider each of these approaches in turn:

2 "Hard-coding" to prevent certain unethical actions

There are two problems that we see with this approach. The first is that, except for the simplest systems, it is impossible to anticipate all the ways such systems could be used or behave unethically. Consider a robot that is designed to be an eldercare assistant. How could one anticipate all the possible circumstances that the robot could find itself in, and make a decision about which actions should be forbidden in each of these circumstances, to build these prohibitions into the behavior of the robot?

The second problem is even if one could succeed in anticipating all the possible unethical actions and block them, these systems would not necessarily perform in the ethically best manner, which is what we should strive for.

3 Putting the burden on the user to ensure ethical behavior by simply providing warnings

Again, this approach requires that we can anticipate all the ways such systems could be used that would be considered to be unethical, to provide proper warnings. Also, not everyone will read the warnings, and take them to heart, and some may even deliberately do that which one is not supposed to do with the program or machine. Finally, this puts too much of a burden on the user. It is preferable for the developers to make sure that such systems can only be used in an ethically acceptable manner.

4 Learning what is ethically acceptable from past decision-makers' judgements or current polls of the general populace and have this guide system behavior

We are now realizing that using earlier decision-makers' judgements to discover values that should govern the behavior of such systems is questionable, because they have revealed biases, e.g., against women and minorities [1]. Isn't this also true of the values learned from current polls? Do we really believe that people in general, when questioned, will give the ethically best answer when asked what one should do in particular situations? Human beings have evolved to favor themselves, their family, and their group. This would surely manifest itself in what they consider to be acceptable behavior, behavior that ethicists would find questionable. And it is doubtful, since people have different loyalties, that they would agree as to what they believe is ethically acceptable behavior. How would such disagreement be handled? Eliminate outliers, going with the answers the largest number of people say is correct? But hasn't history shown repeatedly that the views of someone who was an "outlier" in one period of time turn out to be accepted, even advocated, at a later time? This has been true of knowledge acquisition in every field, and we believe that it is true of ethics as well. The majority once approved of slavery and women being the second-class citizens, but most people today will be reluctant to approve of these views now as a result of "outliers" questioning the views of the majority. Yet, the now "enlightened" ones may not appreciate the extent to which past overt biases are still adversely affecting current practices. Those who want to use objective criteria to fill jobs or for college entrance, thinking that they are advocates of fair procedures, may not realize, for example, that a minority candidate may have had obstacles to face that have resulted in an unequal playing field. Rather than trying to capture the values of even "enlightened" people, we believe that it is important that the values exemplified in such systems have resulted from consulting those with ethical expertise who have examined all of the factors that are ethically relevant.

Finally, those who advocate abstracting values from data derived from earlier decision-makers or the public are just like those who put the burden on the user of their products, avoiding taking responsibility for the values implicit in this data. They need to realize that in areas where human beings' (and perhaps other sentient beings') welfare is at stake, there are always value judgements involved and it is essential that these value judgements are rigorously examined.

5 Using an existing ethical principle or theory to guide system behavior

What is promising about this idea, besides it is having been advocated by at least some ethicists, is that we will build an ethical principle into the program or machine to govern *all* its behavior, rather than have designers just try to block unethical behavior.

The two most widely discussed the existing ethical principles represent two very different approaches to ethics: a consequentialist approach and a deontological approach. Consequentialists believe that the rightness or wrongness of actions depends entirely on the consequences of those actions, whereas those advocating the deontological approach believe that the rightness and wrongness of actions depends upon the nature of those actions in and of themselves, regardless of the consequences.

The most popular consequentialist ethical theory is Act Utilitarianism. Developed by Jeremy Bentham [2] and John Stuart Mill [3], Act Utilitarianism maintains: That action is right which, of all the alternative actions open to the agent, is likely to lead to the great net good consequences, or the least harm, taking all those affected by the action into account. Essentially, as Bentham pointed out, the theory involves performing "moral arithmetic." A machine is certainly capable of doing arithmetic, given the requisite data.

We do not believe, however, that Act Utilitarianism is the ideal theory for a machine to follow. Critics of Act Utilitarianism have pointed out that it can violate human beings' rights, sacrificing one person for the greater net good. It can also conflict with our notion of justice—what people deserve—because the rightness and wrongness of actions is determined entirely by the future consequences of actions, whereas what people deserve is a result of past behavior.

Kant's Categorical Imperative [4], following the deontological approach, focuses on the intrinsic nature of actions, rather than their consequences: One should act in such a manner that one could wish the principle on which one is acting to become a universal law. An essential test, for Kant, was to see whether it is possible for the principle on which one considers acting to be universalized without contradicting itself. Consider the following simple example: one is thinking of reading the newspaper over the shoulder of another person, instead of buying a newspaper oneself, on public transportation into work. The principle on which one is considering acting cannot be universalized without it becoming self-defeating, because no one would have a newspaper!

A secondary test for Kant's Categorical Imperative would have one see if the action would still seem to be acceptable if one puts oneself on the receiving end of the action. Consider this situation that a captain of a ship once faced: The ship has sunk in bad weather, just one lifeboat has survived, and twice as many people as the lifeboat can hold, without it to sinking, are trying to get into it. Naval law requires that the captain makes the decision as to who has a chance to ride out the storm in the lifeboat before, hopefully, help arrives. Suppose he considers adopting the principle that only the strongest people should have places in the lifeboat, because he has no idea what sort of ordeal lies ahead. But now have him put himself in the position of an injured person who will be forced to drown if this principle is followed. It probably would not seem to be acceptable to him. No discriminatory principle would seem to justify letting some die to save others, so everyone should be treated alike; and if so, they will all die as everyone trying to get into the lifeboat will cause it to sink. Aren't the consequences important here? As tragic as the situation is, isn't it better to save some, rather than no one being given a chance to live?

Ross in 1930 [5] advocated combining elements of both consequentialist and deontological reasoning in his approach to ethics. He maintained that the reason why making ethical decisions is so difficult is because it does not involve following a *single* absolute principle, but juggling many duties, some consequentialist, and others deontological, that can be at odds with one another. He maintained that all ethical duties should be considered to be prima facie, which means that although we should attempt to follow them, they each could be overridden in certain situations, when another duty or duties become stronger. His own list of prima facie duties included: the duties of Fidelity, Reparation, Gratitude, Justice, Beneficence, Non-Maleficence, and Self-Improvement. He stated, however, that one could have a different list and we believe that an advantage of Ross' theory is that the list of prima facie duties could vary according to the domain in which the AI developed entity functions.

The major drawback with this approach, however, is that it needs to be supplemented with a decision procedure for cases where the prima facie duties give conflicting advice. Ross, himself, gave us no guidance as to how to solve this problem, but left it up to the agent's intuition which, of course, will not work for the systems under discussion and is not really satisfactory for a human being either, who could rationalize doing whatever he or she wants, finding a duty that could justify that action and maintain that it should be the strongest duty in the current situation.

6 Imposing a hierarchy of ethical duties to guide system behavior

Perhaps what we need to decide cases where there are multiple duties or principles that could conflict with one another is to arrange them in a hierarchy, where the top duty always takes precedence over the second one in the hierarchy, and the second one over the third, etcetera. This is the approach science fiction writer Isaac Asimov considered and first introduced in his 1942 story "Runaround" [6] as laws to govern the behavior of robots:

A robot may not injure a human being, or through inaction, allow a human being to come to harm.

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

These laws were designed to counteract the many science fiction stories where robots were created and destroyed their creators. They seem to ensure the continued domination of humans over robots.

Roger Clarke [7] has pointed out that there are a number of inconsistencies and ambiguities in the laws. What should a robot do when two humans give contradictory instructions, for example? And what exactly counts as "harm" in the first law? It is clear that Asimov himself rejected the laws as a suitable basis for ethics for sophisticated robots. His story *The Bicentennial Man* begins with the laws and the rest of the story provides a refutation of those laws. In one scene in the story, some bullies order the intelligent, creative, and virtuous robot Andrew to destroy himself. He must obey, according to the Laws, because the second law, obeying humans, ranks higher than the third law of self-preservation. Fortunately, another human appears on the scene to counter the bullies' order.

Aside from these problems with Asimov's particular hierarchical laws, what Ross would find objectionable with the hierarchical approach to dealing with multiple laws/duties is the fact that the ones lower on the hierarchy should *never* trump a law/duty that is above.

7 Attempting to learn what is ethically acceptable from those with expertise in ethics

Following this approach, we have worked on developing a general method for representing and resolving ethical issues raised when intelligent, autonomously functioning systems interact with humans in any domain [8]. Using clear cases of ethical dilemmas in a particular domain, where ethicists (using this term as a shorthand for those who have thought long and deeply about ethical matters and can justify their views) are in agreement as to the correct answer, they can be represented in terms of ethically relevant features, and corresponding *prima facie* duties. From this information, machine learning can be used to discover an ethical principle that enables a system to determine the ethically preferable action at each moment.

There must be at least one ethically relevant *feature* in an ethical dilemma that needs to be considered in determining the right action (e.g., that someone could be harmed). Ethically relevant features of dilemmas lead to *prima facie duties* incumbent upon the agent. There is at least one duty, therefore, incumbent upon the agent in an ethical dilemma, either to maximize or minimize the ethical feature(s). Harm, for example, should be minimized. Benefit, on the other hand, should be maximized.

We accept Jeremy Bentham's insight [2] that ethical features, and correlative duties, may be present to a greater or lesser degree in ethical dilemmas. We do not need to specify a *precise* amount of these degrees, needing only to differentiate between degrees that are required to distinguish between ethically distinct situations. For example, a medication that has been prescribed for a patient solely to relieve unpleasant symptoms associated with an illness could be described as involving some benefit for the patient, whereas a medication designed to cure the illness could be described as involving many benefits for the patient. We need these two levels of benefit, because we would likely want to say that a patient refusing to take the first type of medication can be accepted by an eldercare robot, whereas the patient refusing to take the second type is grounds for concern and the robot should notify the doctor.

Since, typically, in every domain where an autonomous system's behavior affects human beings, there are will be a number of prima facie duties to consider that may conflict in ethical dilemmas that are encountered, we can see that we need a decision principle to give us the correct answer when this happens. John Rawls's "reflective equilibrium" approach [9] to creating and refining ethical principles seems reasonable and can be used to solve the problem of coming up with a decision principle when there are several prima facie duties that give conflicting advice in ethical dilemmas. This approach involves generalizing from considered judgments about particular clear cases (that is, where the correct answer seems uncontroversial), testing those generalizations on further cases, and then repeating this process toward the goal of developing a principle that agrees with considered judgments and that can be used to determine the correct action when prima facie duties give conflicting advice. The principle learned evolves as inconsistencies are resolved and new cases are added. Possible actions that can be taken in

a given situation can be represented as sets of degrees of satisfactions or violations of prima facie duties that can be compared to determine which action is ethically preferable according to the learned decision principle.

We believe that attempting to formulate ethics for such systems allows us to have a fresh start at determining the ethical principles that should resolve ethical dilemmas. Because we are concerned with the behavior of autonomous systems, how we think they should treat us, we can be more objective in examining ethics than we would be in discussing human behavior, even though what we come up with should be applicable to human behavior as well.

Furthermore, work in his area, we believe, will very likely bear fruit in the study of ethics in general by discovering principles implicit in the considered judgements of ethicists that have not been stated before and by forcing the examination, leading to a resolution, of inconsistencies revealed through the analysis of cases. Resolution will typically occur through deciding that there is an additional feature present in one case, but not the other, or the range of the intensities of existing features must be expanded.

Finally, perhaps, the most important thing which we can contribute to the welfare of human beings is a vision of how one ought to interact with others. Those working on the ethics of autonomous systems can take a leading role in this worthwhile enterprise by creating ideal role models to inspire human beings to behave more ethically. We can create ethical entities that not only aid us in many ways, but can also show us how we need to behave if we are to survive as a species.

Acknowledgements This material is based in part on work supported by the NSF under Grant Numbers IIS-0500133, IIS-1151305, and IIS-1449155.

Compliance with ethical standards

Conflict of interest None.

Availability of data and material None.

Code availability None.

References

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019)
2. Bentham, J.: An Introduction to the Principles and Morals of Legislation. Oxford Univ Press, Oxford (1799)
3. Mill, J.S.: Utilitarianism. Parker Son and Bourn, London (1863)
4. Kant, I.: Groundwork of the Metaphysic of Morals. Cambridge University Press, Riga (1785)
5. Ross, W.D.: The Right and the Good. Oxford University Press, Oxford (1930)
6. Asimov, I.: Runaround. I Robot. Doubleday, New York City (1950)
7. Clarke, R.: Asimov's laws of robotics. In: Anderson, M., Anderson, S. (eds.) Machine Ethics, pp. 254–284. Cambridge University Press, Cambridge (2011) https://doi.org/10.1017/CBO9780511978036.020
8. Anderson, M., Anderson, S.: GenEth: a general ethical dilemma analyzer. Paladyn J. Behav. Robot. **9**(1), 337–357 (2018). https://doi.org/10.1515/pjbr-2018-0024
9. Rawls, J.: Outline of a decision procedure for ethics. Philos. Rev. **60**(2), 177–197 (1951)