The Relationship Between Intelligent, Autonomously Functioning Machines and Ethics

Susan Leigh Anderson
University of Connecticut

Michael Anderson
University of Hartford

Abstract.

In this paper we argue that intelligent, autonomously functioning machines whose behaviour affects human beings should be considered to be moral agents, although they cannot be held morally responsible for their actions. We believe that work on machine ethics – training intelligent, autonomously functioning machines to act in an ethically responsible manner – is essential for the public to feel comfortable interacting with the machines that are currently being created and, furthermore, will enable us to consider developing other machines that can provide benefits for humans. Giving such machines general ethical principles to follow that could cover any possible set of circumstances, even those that were unanticipated, would be ideal. This approach offers the added benefits that the principles could be specified to the user and overseer to justify the behavior of the machine and the principles could be modified if deemed advisable. Developing principles for this use is a complex process and new tools and methodologies will be needed to help contend with this complexity. We are in the process of developing such a tool and methodology.

1 INTRODUCTION

Could machines ever be considered to be moral agents? If so, how can we ensure that they behave in an ethically responsible manner? How does moral agency relate to moral responsibility? Could intelligent, autonomously functioning machines be viewed as moral agents that are not morally responsible for their actions? Might *we* have a moral responsibility to (a) develop ethically trained machines that can bring about desirable states of affairs and (b) harness machine capabilities to further our understanding of ethics? These are questions that we will consider first as we explore the relationship between intelligent machines and ethics. We will then briefly summarize our previous research in machine ethics, as well as our current research that enables a system to learn the ethically relevant features (and required range of satisfaction/violation) of ethical dilemmas, corresponding *prima facie* duties, and decision principles to resolve conflicts between those duties.

2 MORAL AGENTS

Could machines ever be considered to be *moral agents*? To answer this question, we should consider James Moor's five categories [1] of ways in which values might be ascribed to machines. The first three clearly fall short of being true moral agenthood: *normative agents*, machines designed with a specific purpose in mind, e.g. a proof checker, that are only assessed as to how well they satisfy that purpose; *ethical impact agents*, that are not only designed with a specific purpose, but also have an, ideally positive, impact on the world such as the robot jockeys that guide camels in races in Qatar, replacing young boys who are thereby freed from slavery; and *implicit ethical agents*, machines that have been programmed by human designers to behave in ways that are consistent with ethical practices.

What is necessary in order to be a *moral agent* is that the agent "considers" options for possible actions that it could perform, selects the one that, following some ethical theory, would be the most ethically correct one and performs that action. The last two of Moor's categories both satisfy these requirements: *explicit ethical agents*, machines that calculate the best action when faced with ethical

dilemmas by being able to represent the situation they are faced with, "consider" which actions are possible in the situation, assess those actions in terms of an ethical theory, and then perform the action that they have determined to be the most ethically correct one; and *full ethical agents*, a term which requires, in addition to being an explicit ethical agent, that the agent acts intentionally, is conscious and has free will, thus enabling the agent to be held morally responsible for its actions.

Let us assume here what is commonly accepted, that human actions satisfy these criteria while machine actions do not, and very likely never will. If so, then *ethical agenthood should not be equated with being held morally responsible for one's actions*, but rather, the second is a subset of the first. It is possible to be an (explicit) ethical agent and yet one should not be held morally responsible for one's actions. An ethically trained, autonomously functioning machine is a prime example.

3 CREATING AN ETHICAL MACHINE

How can we ensure that intelligent, autonomously functioning machines behave in an ethically acceptable fashion? First, it's important to recognize, as roboticists often don't, that there are ethical ramifications of *all* machine behavior that affects humans. This is why ethicists need to be involved in their development.

If a machine is to function autonomously, it may be very difficult to anticipate each situation of concern that may arise and ensure that the machine will act in an ethically acceptable manner in that situation. It would be far better to program the machine with general ethical principles that could cover any possible set of circumstances, even those that were unanticipated, with the added benefits that the principles could be given to the user and overseer to justify the behavior of the machine and the principles could be modified if deemed advisable.

We believe that work on machine ethics – training intelligent, autonomously functioning machines to act in an ethically responsible manner – is not only essential for the public to feel comfortable interacting with the machines that are currently being created, but will also enable us to consider developing other machines that can provide benefits for humans. Indeed, we believe that we have an ethical obligation to do so. Correct ethical behavior does not only involve *not* doing certain things, but also *attempting to bring about ideal states of affairs*. If a robotic personal assistant could be developed that acts in an ethically responsible manner and allows an elderly person who wishes to live at home alone to safely do so, notifying a doctor or relative when its charge needs attention, this would clearly be a good thing to do. What other machines could be developed, with ethical principles guiding their behavior, that could improve our lives? There should, of course, be a veto to some machine development proposals from an ethical perspective: If ethicists disagree as to the ethical principles that are needed to govern the behavior of certain types of machines, they should not be produced.

We believe that work on machine ethics is likely to advance the study of ethics, ideally leading to there being fewer disagreements as to what counts as ethically correct behavior for both machines and humans. The ethics that we consider embodying in a machine will, of necessity, have to be sharpened to a degree that ethics has never been sharpened before, because machines cannot be programmed in a vague manner. Machine ethics is concerned with the application of ethical theory to specific domains in which machines could function, forcing scrutiny of the details involved in actually applying ethical principles to particular real-life cases, rather than the artificial examples that ethicists typically discuss. We can't help but learn something about ethics from this endeavor.

We can even harness machine intelligence to discover the general ethical principles that we would like machines to follow, using inductive reasoning, by generalizing from information given to them about the desired behavior in particular cases. What computers are good at is keeping track of lots of information that quickly overwhelms humans; and formal representation of ethical dilemmas and their solutions make it possible for machines to spot contradictions that need to be resolved.

Furthermore, we believe that machine ethics research allows us to have a fresh perspective on ethics, very much like John Rawls' thought experiment for determining the principles of justice. [2] Just as he suggested that we adopt a "veil of ignorance" perspective, where we do not know our positions in life, we will consider how we would like machines to treat us, instead of trying to rationalize what we can get away with doing so as to protect our own positions in life.

Finally, since the machine ethics research community is international, our hope is that we will one day be able to come up with a set of ethical principles that rational persons world-wide can accept. Embodying these principles in machines will give us good role models for how *we* ought to behave, perhaps leading to less unethical human behavior, giving us a better chance of being able to survive as a species.

4 OUR PREVIOUS RESEACH

Our own research has proceeded in a step-wise fashion, each time doing the task we set ourselves at a proof of concept level. We [3], first, created the program "Jeremy", an attempt to capture the thinking of a well-known ethical theory, Hedonistic Act Utilitarianism [4], that maintains that ethics is a matter of doing "moral arithmetic". We considered two possible actions that could be performed in ethical dilemmas, whether those who would be affected were likely to receive pleasure or displeasure from each of the actions being done and, if so, whether it was just some, or considerable, pleasure or displeasure. Creating this program provided a good starting point in attempting to make ethics computable, and we learned a lot from creating "Jeremy", but we never considered it to be the ideal ethical theory for machine or human ethics. We believe that the correct approach to ethics is a theory that combines elements of deontological and utilitarian thinking, one that can take into account justice considerations, in addition to the likely future consequences of possible actions that could be performed.

The *prima facie duty* approach to ethics, which we owe to W.D. Ross [5], is ideal for combining multiple ethical obligations and can be adapted to many different domains by simply changing the various prima facie duties. There is one serious drawback with this approach, however, where there are a number of ethical duties that we should try to follow, each of which can be overridden on occasion by one of the other duties: There is no decision principle for determining which duty should prevail when the prima facie duties pull in different directions.

The next task we set ourselves, therefore, was to see if we could harness machine capabilities to discover a decision principle for a prima facie duty approach to ethics. Since we were looking for a prototype solution to the problem, we constrained the task. We used a well-known prima facie duty theory in the domain of biomedicine with a limited number of duties and applied it to a common, but narrow, type of ethical dilemma in that domain to develop and test our solution to the problem.

The prima facie duty theory that we used is Beauchamp and Childress' Principles (Duties) of Biomedical Ethics. [6] The type of dilemma that we considered [7] involved three of their four duties: Respect for the Autonomy of the patient, Nonmaleficence (not causing harm to the patient) and Beneficence (promoting patient welfare). The general type of ethical dilemma that we considered was: A health care professional has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment option. Should the health care worker try again to change the patient's mind or accept the patient's decision as final? Besides the duty to respect patient autonomy, this type of dilemma involves the duty not to cause harm to the patient (nonmaleficence) and/or the duty to promote patient welfare (beneficence), since the recommended treatment is designed to prevent harm to, and/or benefit, the patient.

The options for the health care professional are just two – either to accept the patient's decision or not – and there are a finite number of specific types of cases using the representation scheme we adopted for possible cases. Our representation scheme consisted of an ordered set of values for each of the possible actions that could be performed, where those values reflected whether the duties were satisfied or violated (if they were involved) and, if so, to which of two possible degrees. We learned

from Bentham, in our earlier work, that the degree of satisfaction or violation of a duty can be very important. It turns out that, with our allowable range of values for the three possible duties that could be at stake, there are 18 possible case profiles.

Inspired by John Rawls' "reflective equilibrium" [8] approach to creating and refining ethical principles, we used inductive logic programming (ILP) to discover a decision principle from being given the best action in just 4 cases that correctly covered 14 of the remaining 18 possible cases. The principle learned was, of course, implicit in the judgments that ethicists provided concerning the 4 cases; but, to our knowledge, it had never been stated before.[1] It gives us hope that not only can ethics help to guide machine behavior, but that machines can help us to discover the ethics needed to guide such behavior. Furthermore, we developed a way of representing the needed data and a system architecture for implementing the principle.

We then went on to develop three applications of the principle: (1) MedEthEx [7], an interactive medical ethics advisor system for dilemmas of the type that we considered; (2) EthEl [9], a medication reminder system that used the learned principle to not only issue reminders at appropriate times, but also determined when an overseer should be notified if the patient refuses to take the medication. Finally, (3) we instantiated EthEl in a Nao robot [10], the first example, we believe, of a robot that follows an ethical principle in determining which actions it will take. Nao is capable of finding and walking towards a patient who needs to be reminded to take a medication, bringing the medication to the patient, engaging in a natural language exchange, and notifying an overseer by e-mail when necessary.

6 OUR CURRENT RESEARCH

In constraining the task for discovering a decision principle in a particular limited domain, we previously made a number of assumptions, most notably using a particular set of prima facie duties and a particular range of possible satisfaction or violation of those duties. The current challenge we are working on is to develop a method for generating from scratch, through an automated interactive dialogue with an ethicist, the ethics needed for a machine to function ethically in a particular domain, without making the assumptions of particular prima facie duties and range of intensity used in our earlier decision principle learning prototype. We now see that what is most basic to ethical dilemmas is that there is at least one feature of an ethical dilemma that makes it of ethical concern (e.g. that someone could be harmed) and there must be at least one ethical duty incumbent upon the agent to either maximize or minimize that feature (e.g. harm should be minimized). Features, duties, range of duty satisfaction or violation, and needed decision principles will be systematically learned by the machine through automated interaction with ethicists, using examples of dilemmas provided by ethicists.[2]

The introduction of new features, corresponding duties and a wider range of duty satisfaction/violation are generated through resolving contradictions that arise as new cases are introduced. With two ethically identical cases – i.e. cases with the same ethically relevant feature(s) to the same degree – an action cannot be right in one of the cases, while the comparable action in the other case is considered to be wrong. Formal representation of ethical dilemmas and their solutions make it possible for machines to spot contradictions that need to be resolved. A contradiction may arise when trying to represent a new case using existing duties and ranges, if the opposite action is deemed preferable when compared with an earlier case with the same profile. If both judgments are correct, there must be either a *qualitative* distinction between them (which requires a new feature and duty) or a *quantitative* distinction (which requires that the range of existing duties must be expanded).

---

[1] For a summary of the prima facie duties used and type of ethical dilemma considered, as well as the principle that was learned, see [11].

[2] See [11] for preliminary work on this new approach.

Imagining a dialogue between the learning system and an applied ethicist, using our medication reminder system as an example, we can see that (in principle) we can hone down what is required to enable the ethicist to begin to teach the system the ethically relevant features, correlative duties and eventually the range of intensities required, from which decision principles can be discovered. The system prompts the ethicist to give an example of an ethical dilemma that a medication reminder system might face, asking the ethicist to state the possible actions that could be performed, which one is preferable, and what feature is present in one of the actions, but not in the other. From this information, a duty that is at least *prima facie* can be inferred, either to maximize or minimize the feature, depending upon whether the action that has the feature is preferable or not. Information is stored in the system, including a representation of a *positive* case (that one action is preferable to the other) and a *negative* one (that the opposite action is not preferable).

The system might then prompt the ethicist to give an example of a new ethical dilemma where the judgment of the ethicist would be the reverse of the first case (i.e. instead of notifying the overseer as being correct, one should not notify the overseer). Prompting the ethicist, the system determines whether in this case a *second* feature is present, which should be maximized or minimized, or whether the difference between the two cases amounts to a difference in the *degree* to which the original feature is present. As new features are introduced, with corresponding prima facie duties, and ranges of intensity, the system begins to formulate and then refine a decision principle to resolve cases where the prima facie duties pull in different directions. We envision the system prompting the ethicist to enter in just the types of cases that will enable it to obtain the data it needs to learn a decision principle as efficiently as possible, i.e. to infer an ethically acceptable decision principle with the fewest number of cases.

There are two advantages to discovering ethically relevant features/duties, and an appropriate range of intensities, with this approach to learning what is needed to resolve ethical dilemmas. First, it can be tailored to the domain with which one is concerned. Different sets of ethically relevant features/prima facie duties can be discovered, through considering examples of dilemmas in the different domains in which machines will operate. A second advantage is that features/duties can be added or removed, if it becomes clear that they are needed or redundant.

In addition, we believe that there is hope for discovering decision principles that, at best, have only been implicit in the judgments of ethicists and may lead to surprising new insights, and therefore breakthroughs, in ethical theory. This can happen as a result of the computational power of today's machines that can keep track of more information than a human mind and require consistency. Inconsistencies that are revealed will force ethicists to try to resolve those inconsistencies through the sharpening of distinctions between ethical dilemmas that appear to be similar at first glance, but which we want to treat differently. There is, of course, always the possibility that genuine disagreement between ethicists will be revealed concerning what is correct behavior in ethical dilemmas in certain domains. If so, the nature of the disagreement should be sharpened as a result of this procedure; and we should not permit machines to make decisions in these domains.

More formally, we represent a possible *action* that could be performed as a tuple of satisfaction/violation values of the duties corresponding to the ethically relevant features the action involves, where an *ethically relevant feature* is a detail used to determine whether an action is right or wrong and a *duty* is the minimization or maximization of an ethically relevant feature. An *ethical dilemma* is the tuple of differentials of the satisfaction/violation values of the corresponding duties two actions involve. Finally, a decision *principle* defines a transitive binary relation over a set of actions that partitions it into subsets ordered by ethical preference with actions within the same partition having equal preference.

While we believe that the representation scheme that we have been developing will be helpful in categorizing and resolving ethical dilemmas in a manner that permits machines to behave more ethically, we envision an extension and an even more subtle representation of ethical dilemmas in future research. We need to consider more possible actions available to the agent, where there is not necessarily a symmetry between actions (i.e. where the degree of satisfaction/violation of a duty in

one is mirrored by the opposite in the other). Also, ideally, one should not only consider present options, but possible actions that could be taken in the future. It might be the case, for instance, that one present option, which in and of itself appears to be more ethically correct than another option, could be postponed and performed at some time in the future, whereas the other one cannot, and this should affect the assessment of the actions.

7 CONCLUSION

Intelligent, autonomously functioning machines whose actions affect human beings should be considered to be moral agents. Thus it can be argued that *machine ethics* ought to be the driving force in determining the extent to which they should be allowed to function. Autonomous systems that behave in a less than ethically acceptable manner towards human beings will not, and should not, be tolerated. Thus, it becomes paramount that we demonstrate that these systems will not violate the rights of human beings and will perform only those actions that follow acceptable ethical principles. Principles offer the further benefits of serving as a basis for justification of actions taken by a system as well as for an overarching control mechanism to manage unanticipated behavior of such systems. Developing principles for this use is a complex process and new tools and methodologies will be needed to help contend with this complexity. We are in the process of developing such a tool and methodology.

ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers IIS-0500133 and IIS-1151305.

REFERENCES

[1] Moor, J., "The Nature, Importance, and Difficulty of Machine Ethics," in *Machine Ethics* (Anderson, M. and Anderson, S., eds.), Cambridge University Press, New York, NY, pp. 13-20, 2011.

[2] Rawls, J., *A Theory of Justice*, Harvard University Press, 1971.

[3] Anderson, M., Anderson, S., and Armen, C., "An Approach to Computing Ethics," *IEEE Intelligent Systems*, Vol. 21, no. 4, 2006.

[4] Bentham, J., *An Introduction to the Principles of Morals and Legislation*, Chapter 17 (Burns, J. and Hart, H., eds.) Clarendon Press, Oxford, 1969.

[5] Ross, W.D., *The Right and the Good*, Oxford University Press, Oxford, 1930.

[6] Beauchamp and Childress, *Principles of Biomedical Ethics*. Oxford, UK: Oxford University Press, 1979.

[7] Anderson, M., Anderson, S. and Armen, C., "MedEthEx: A Prototype Medical Ethics Advisor" in *Proceedings of the Eighteenth Conference on Innovative Applications of Artificial Intelligence*, Boston, Massachusetts, August (2006b).

[8] Rawls, J., "Outline for a Decision Procedure for Ethics", *The Philosophical Review* 60(2): 177–197, 1951.

[9] Anderson, M. and Anderson, S., "EthEl: Toward a Principled Ethical Eldercare Robot" in *Proceedings of the AAAI Fall 2008 Symposium on AI in Eldercare: New Solutions to Old Problems*, Arlington, Virginia, November (2008).

[10] Anderson, M. and Anderson, S., "An Ethical Robot", *Scientific American*, October (2010).

[11] Anderson, S. and Anderson, M., "A Prima Facie Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists," in *Machine Ethics* (Anderson, M. and Anderson, S., eds.) Cambridge University Press, New York, NY, pp. 476-492 (2011).