2

Machine Metaethics

Susan Leigh Anderson

The newly emerging field of *machine ethics* is concerned with ensuring that the behavior of machines toward human users is ethically acceptable. There are domains in which intelligent machines could play a significant role in improving our quality of life as long as concerns about their behavior can be overcome by ensuring that they behave ethically. *Machine metaethics* examines the field of machine ethics. It talks *about* the field, rather than doing work in it. Examples of questions that fall within machine metaethics are: How central are ethical considerations to the development of artificially intelligent agents? What is the ultimate goal of machine ethics? What does it mean to add an ethical dimension to machines? Is ethics computable? Is there a single correct ethical theory that we should try to implement? Should we expect the ethical theory we implement to be complete? That is, should we expect it to tell a machine how to act in every ethical dilemma? How important is consistency? If it is to act in an ethical manner, is it necessary to determine the moral status of the machine itself?

When does machine behavior have ethical import? How should a machine behave in a situation in which its behavior does have ethical import? Consideration of these questions should be central to the development of artificially intelligent agents that interact with humans. We should not be making intelligent machines unless we are confident that they have been designed to "consider" the ethical ramifications of their behavior and will behave in an ethically acceptable manner. Furthermore, in contemplating designing intelligent machines, ethical concerns should not be restricted to just prohibiting unethical behavior on the part of machines. Rather, they should extend to considering the additional tasks that machines could perform given appropriate ethical guidance and, perhaps, also to considering whether we have an obligation to develop ethical intelligent machines that could enhance human lives. Just as human ethics is concerned both with what we *ought not* to do and what we *ought* to do – it is unethical for people to cheat others and ethically praiseworthy for people to help others during a crisis, for example – so we should be thinking both about ensuring that machines *do not*

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press

do certain things and about creating machines that *do* provide benefits to humans that they would otherwise not receive.

The ultimate goal of machine ethics, I believe, is to create a machine that follows an ideal ethical principle or a set of ethical principles in guiding its behavior; in other words, it is guided by this principle, or these principles, in the decisions it makes about possible courses of action it could take. We can say, more simply, that this involves "adding an ethical dimension" to the machine.

It might be thought that adding an ethical dimension to a machine is ambiguous. It could mean either: (a) designing the machine with built-in limitations to its behavior or requiring particular behavior according to an ideal ethical principle or principles that are *followed by the human designer*; or (b) giving *the machine* (an) ideal ethical principle(s) or some examples of ethical dilemmas together with correct answers, and a learning procedure from which it can abstract (an) ideal ethical principle(s), so that *it* can use the principle(s) in guiding its own actions. In the first case, it is the human being who is following ethical principles and concerned about harm that could come from machine behavior. This falls within the well-established domain of what has sometimes been called "computer ethics," rather than machine ethics. In the second case, however, the machine itself is reasoning on ethical matters, which is the ultimate goal of *machine* ethics.[1] An indication that this approach has been adopted can be seen if the machine can make a judgment in an ethical dilemma with which it has not previously been presented.

In order for it to be accepted as ethical by the human beings with whom it interacts, it is essential that the machine has an ethical principle or a set of principles that it uses to calculate how it ought to behave in an ethical dilemma, because it must be able to *justify* its behavior to any human being who has concerns about its actions. The principle(s) it uses to calculate how it should behave and justify its actions, furthermore, must be translatable into ordinary language that humans can understand and must, on reflection, appear to be intuitively correct. If the machine is not able to justify its behavior by giving (an) intuitively correct, understandable ethical principle(s) that it has used to determine its actions, humans will distrust its ability to consistently behave in an ethical fashion.

Central to the machine ethics project is the belief (or hope) that ethics can be made computable, that it can be sharpened enough to be able to be programmed into a machine. Some people working on machine ethics have started tackling the challenge of making ethics computable by creating programs that enable machines to act as ethical advisors to human beings, believing that this is a good first step toward the eventual goal of developing machines that can follow ethical principles in guiding their own behavior (Anderson, Anderson, and Armen 2005).[2]

[1] Also, only in this second case can we say that the machine is functioning autonomously.

[2] Bruce McLaren has also created a program that enables a machine to act as an ethical advisor to human beings, but in his program the machine does not make ethical decisions itself. His advisor system simply informs the human user of the ethical dimensions of the dilemma without reaching a decision (McLaren 2003).

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press

Four pragmatic reasons could be given for beginning this way: (1) One could start by designing an advisor that gives guidance to a select group of persons in a finite number of circumstances, thus reducing the scope of the assignment.[3] (2) Machines that just advise human beings would probably be more easily accepted by the general public than machines that try to behave ethically themselves. In the first case, it is human beings who will make ethical decisions by deciding whether to follow the recommendations of the machine, preserving the idea that *only human beings will be moral agents*. The next step in the machine ethics project is likely to be more contentious: creating *machines that are autonomous moral agents*. (3) A big problem for Artificial Intelligence in general, and so for this project too, is how to get needed data, in this case the information from which ethical judgments can be made. With an ethical advisor, human beings can be prompted to supply the needed data. (4) Ethical theory has not advanced to the point where there is agreement, even by ethical experts, on the correct answer for all ethical dilemmas. An advisor can recognize this fact, passing difficult decisions that have to be made in order to act to the human user. An autonomous machine that is expected to be moral, on the other hand, would either not be able to act in such a situation or would decide arbitrarily. Both solutions seem unsatisfactory.

This last reason is a cause for concern for the entire machine ethics project. It might be thought that for ethics to be computable, we must have a theory that determines which action is morally right in every ethical dilemma. There are two parts to this view: (1) We must know which is the correct ethical theory, according to which the computations are made; and (2) this theory must be *complete*, that is, it must tell us how to act in any ethical dilemma that might be encountered.

One could try to avoid making a judgment about which is the correct ethical theory (rejecting 1) by simply trying to implement *any* ethical theory that has been proposed (e.g., Hedonistic Act Utilitarianism or Kant's Categorical Imperative), making no claim that it is necessarily the *best* theory and therefore the one that ought to be followed. Machine ethics then becomes just an exercise in what can be computed. However, this is surely not particularly worthwhile, unless one is trying to figure out an approach to programming ethics in general by practicing on the theory that is chosen.

Ultimately one has to decide that a particular ethical theory, or at least an approach to ethical theory, is correct. Like W. D. Ross (1930), I believe that the simple, single absolute duty theories that have been proposed are all deficient.[4] Ethics is more complicated than that, which is why it is easy to devise a counterexample to any of these theories. There are advantages to the multiple *prima facie* duties[5] approach that Ross adopted, which better captures conflicts that often

[3] This is the reason why Anderson, Anderson, and Armen started with "MedEthEx," which advises health care workers – and, initially, in just one particular circumstance.

[4] I am assuming that one will adopt the action-based approach to ethics, because we are concerned with the *behavior* of machines.

[5] A prima facie duty is something that one ought to do unless it conflicts with a stronger duty, so there can be exceptions, unlike an *absolute* duty, for which there are no exceptions.

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press

arise in ethical decision making: (1) There can be different sets of prima facie duties for different domains, because there are different ethical concerns in such areas as biomedicine, law, sports, and business, for example. (2) The duties can be amended, and new duties added if needed, to explain the intuitions of ethical experts about particular cases as they arise. Of course, the main problem with the multiple prima facie duties approach is that there is no decision procedure when the duties conflict, which often happens. It seems possible, though, that a decision procedure could be learned by generalizing from intuitions about correct answers in particular cases.

Does the ethical theory or approach to ethical theory that is chosen have to be complete? Should those working on machine ethics expect this to be the case? My answer is: probably not. The implementation of ethics cannot be more complete than is accepted ethical theory. Completeness is an ideal for which to strive, but it may not be possible at this time. There are still a number of ethical dilemmas in which even experts are not in agreement as to what is the right action.[6]

Many nonethicists believe that this admission offers support for the metaethical theory known as Ethical Relativism. Ethical Relativism is the view that when there is disagreement over whether a particular action is right or wrong, both sides are correct. According to this view, there is no single correct ethical theory. Ethics is relative to either individuals (subjectivism) or societies (cultural relativism). Most ethicists reject this view because it entails that we cannot criticize the actions of others, no matter how heinous. We also cannot say that some people are more moral than others or speak of moral improvement – for example, that the United States has become a more ethical society by granting rights first to women and then to African Americans.

There certainly do seem to be actions that ethical experts (and most of us) believe are absolutely wrong (e.g., slavery and torturing a baby are wrong). Ethicists are comfortable with the idea that one may not have definitive answers for *all* ethical dilemmas at the present time, and even that we may in the future decide to reject some of the views we now hold. Most ethicists believe, however, that *in principle* there are correct answers to all *ethical* dilemmas,[7] as opposed to questions that are just matters of taste (deciding which shirt to wear, for example). Someone working in the area of machine ethics, then, would be wise to allow for gray areas in which one should not necessarily expect answers at this time and even allow for the possibility that parts of the theory being implemented may need to be revised. Care should be taken to ensure that we do not permit

[6] Some who are more pessimistic than I am would say that there will always be some dilemmas about which even experts will disagree as to what is the correct answer. Even if this turns out to be the case, the agreement that surely exists on many dilemmas will allow us to reject a completely relativistic position, and we can restrict the development of machines to areas where there is general agreement as to what is acceptable behavior.

[7] The pessimists would perhaps say: "There are correct answers to many (or most) *ethical* dilemmas."

machines to function autonomously in domains in which there is controversy concerning what is correct behavior.

There are two related mitigating factors that allow me to believe that there is enough agreement on ethical matters that at least some ethical intelligent machines can be created: First, as just pointed out, although there may not be a universally accepted *general theory* of ethics at this time, there is wide agreement on what is ethically permissible and what is not in *particular cases*. Much can be learned from those cases. Many approaches to capturing ethics for a machine involve a machine learning from particular cases of acceptable and unacceptable behavior. Formal representation of particular ethical dilemmas and their solutions make it possible for machines to store information about a large number of cases in a fashion that permits automated analysis. From this information, general ethical principles may emerge.

Second, machines are typically created to function in specific, limited domains. Determining what is and is not ethically acceptable in a specific domain is a less daunting task than trying to devise a general theory of ethical and unethical behavior, which is what ethical theorists attempt to do. Furthermore, it might just be possible that in-depth consideration of the ethics of limited domains could lead to generalizations that could be applied to other domains as well, which is an extension of the first point. Those working on machine ethics, because of its practical nature, have to consider and resolve all the details involved in actually applying a particular ethical principle (or principles) or approach to capturing/simulating ethical behavior, unlike ethical theoreticians who typically discuss hypothetical cases. There is reason to believe that the “real-world” perspective of AI researchers, working with applied ethicists, stands a chance of getting closer to capturing what counts as ethical behavior than the abstract reasoning of most ethical theorists. As Daniel Dennett recently said, “AI makes Philosophy honest” (Dennett 2006).

Consistency (that one should not contradict oneself), however, is crucial, because it is essential to rationality. Any inconsistency that arises should be cause for concern and for rethinking either the theory itself or the way that it is implemented. One cannot emphasize the importance of consistency enough, and machine implementation of an ethical theory may be far superior to the average human being’s attempt at following the theory. A machine is capable of rigorously following a logically consistent principle or set of principles, whereas most human beings easily abandon principles and the requirement of consistency that is the hallmark of rationality because they get carried away by their emotions. Human beings could benefit from interacting with a machine that spells out the consequences of consistently following particular ethical principles.

Let us return now to the question of whether it is a good idea to try to create an ethical advisor before attempting to create a machine that behaves ethically itself. An even better reason than the pragmatic ones given earlier can be given for the

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press

field of machine ethics to proceed in this manner: One does not have to make a judgment about the status of the machine itself if it is just acting as an advisor to human beings, whereas one does have to make such a judgment if the machine is given moral principles to follow in guiding its own behavior. Because of the particular difficulty involved,[8] it would be wise to begin with a project that does not require such judgments. Let me explain.

If the machine is simply advising human beings as to how to act in ethical dilemmas, where such dilemmas involve the proper treatment of other human beings (as is the case with classical ethical dilemmas), it is assumed that either (1) the advisor will be concerned with ethical dilemmas that only involve human beings, or (2) only human beings have moral standing and need to be taken into account. Of course, one *could* build in assumptions and principles that maintain that other beings and entities should have moral standing and be taken into account as well; the advisor could then consider dilemmas involving animals and other entities that might be thought to have moral standing. Such a purview would, however, go beyond universally accepted moral theory and would certainly not, at the present time, be expected of an ethical advisor for human beings facing traditional moral dilemmas.

On the other hand, if the machine is given principles to follow to guide its own behavior, an assumption must be made about its status. This is because in following any ethical theory, it is generally assumed that the agent has moral standing, and therefore he/she/it must consider at least him/her/itself, and typically others as well, in deciding how to act.[9] A machine agent must "know" if it is to count, or whether it must always defer to others who count while it does not, in calculating the correct action in an ethical dilemma.

I have argued that, for many reasons, it is a good idea to begin to make ethics computable by creating a program that would enable a machine to act as an ethical advisor to human beings facing traditional ethical dilemmas. The ultimate goal of machine ethics – to create autonomous ethical machines – will be a far more challenging task. In particular, it will require that a difficult judgment be made about the status of the machine itself. I have also argued that the principle(s) followed by an ethical machine must be consistent, but should not necessarily completely cover every ethical dilemma that machines could conceivably face. As a result, the development of machines that function autonomously must keep pace with those areas in which there is general agreement as to what is considered to be correct ethical behavior. Seen in this light, work in the field of machine ethics should be seen as central to the development of autonomous machines.

[8] See S. L. Anderson, "The Unacceptability of Asimov's 'Three Laws of Robotics' as a Basis for Machine Ethics," included in this volume, which demonstrates how difficult it would be.

[9] If Ethical Egoism is accepted as a plausible ethical theory, then the agent only needs to take him/her/itself into account, whereas all other ethical theories consider others as well as the agent, assuming that the agent has moral status.

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press

References

Anderson M., Anderson S. L., and Armen, C. (2005), "MedEthEx: Towards a Medical Ethics Advisor," in *Proceedings of the AAAI Fall Symposium on Caring Machines: AI and Eldercare*, Menlo Park, California.

Dennett, D. (2006), "Computers as Prostheses for the Imagination," invited talk presented at the International Computers and Philosophy Conference, Laval, France, May 3.

McLaren, B. M. (2003), "Extensionally Defining Principles and Cases in Ethics: An AI Model," in *Artificial Intelligence Journal*, 150 (1–2): 145–1813.

Ross, W. D. (1930), *The Right and the Good*, Oxford University Press, Oxford.

https://doi.org/10.1017/CBO9780511978036.004 Published online by Cambridge University Press