1

The Nature, Importance, and Difficulty of Machine Ethics

James H. Moor

Implementations of machine ethics might be possible in situations ranging from maintaining hospital records to overseeing disaster relief. But what is machine ethics, and how good can it be?

The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions.

A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. You're unlikely to find easily testable hypotheses in the murky waters of philosophy. But we can't – and shouldn't – avoid consideration of machine ethics in today's technological world.

As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged – or will soon engage – in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying.

Varieties of Machine Ethics

When people speak of technology and values, they're often thinking of ethical values. But not all values are ethical. For example, practical, economic, and

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

aesthetic values don't necessarily draw on ethical considerations. A product of technology, such as a new sailboat, might be practically durable, economically expensive, and aesthetically pleasing, absent consideration of any ethical values. We routinely evaluate technology from these nonethical normative viewpoints. Tool makers and users regularly evaluate how well tools accomplish the purposes for which they were designed. With technology, all of us – ethicists and engineers included – are involved in evaluation processes requiring the selection and application of standards. In none of our professional activities can we retreat to a world of pure facts, devoid of subjective normative assessment.

By its nature, computing technology is normative. We expect programs, when executed, to proceed toward some objective – for example, to correctly compute our income taxes or keep an airplane on course. Their intended purpose serves as a norm for evaluation – that is, we assess how well the computer program calculates the tax or guides the airplane. Viewing computers as technological agents is reasonable because they do jobs on our behalf. They're normative agents in the limited sense that we can assess their performance in terms of how well they do their assigned jobs.

After we've worked with a technology for a while, the norms become second nature. But even after they've become widely accepted as the way of doing the activity properly, we can have moments of realization and see a need to establish different kinds of norms. For instance, in the early days of computing, using double digits to designate years was the standard and worked well. But, when the year 2000 approached, programmers realized that this norm needed reassessment. Or consider a distinction involving AI. In a November 1999 correspondence between Herbert Simon and Jacques Berleur,[1] Berleur was asking Simon for his reflections on the 1956 Dartmouth Summer Research Project on Artificial Intelligence, which Simon attended. Simon expressed some puzzlement as to why Trenchard More, a conference attendee, had so strongly emphasized modal logics in his thesis. Simon thought about it and then wrote back to Berleur,

> My reply to you last evening left my mind nagged by the question of why Trench Moore [*sic*], in his thesis, placed so much emphasis on modal logics. The answer, which I thought might interest you, came to me when I awoke this morning. Viewed from a computing standpoint (that is, discovery of proofs rather than verification), a standard logic is an indeterminate algorithm: it tells you what you MAY legally do, but not what you OUGHT to do to find a proof. Moore [*sic*] viewed his task as building a modal logic of "oughts" – a strategy for search – on top of the standard logic of verification.

Simon was articulating what he already knew as one of the designers of the Logic Theorist, an early AI program. A theorem prover must not only generate a list of well-formed formulas but must also find a sequence of well-formed formulas constituting a proof. So, we need a procedure for doing this. *Modal logic* distinguishes between what's permitted and what's required. Of course, both are norms for the subject matter. But norms can have different levels of obligation,

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

as Simon stresses through capitalization. Moreover, the norms he's suggesting aren't ethical norms. A typical theorem prover is a normative agent but not an ethical one.

Ethical-Impact Agents

You can evaluate computing technology in terms of not only design norms (that is, whether it's doing its job appropriately) but also ethical norms.

For example, *Wired* magazine reported an interesting example of applied computer technology.[2] Qatar is an oil-rich country in the Persian Gulf that's friendly to and influenced by the West while remaining steeped in Islamic tradition. In Qatar, these cultural traditions sometimes mix without incident – for example, women may wear Western clothing or a full veil. And sometimes the cultures conflict, as illustrated by camel racing, a pastime of the region's rich for centuries. Camel jockeys must be light – the lighter the jockey, the faster the camel. Camel owners enslave very young boys from poorer countries to ride the camels. Owners have historically mistreated the young slaves, including limiting their food to keep them lightweight. The United Nations and the US State Department have objected to this human trafficking, leaving Qatar vulnerable to economic sanctions.

The machine solution has been to develop robotic camel jockeys. The camel jockeys are about two feet high and weigh 35 pounds. The robotic jockey's right hand handles the whip, and its left handles the reins. It runs Linux, communicates at 2.4 GHz, and has a GPS-enabled camel-heart-rate monitor. As *Wired* explained it, "Every robot camel jockey bopping along on its improbable mount means one Sudanese boy freed from slavery and sent home." Although this eliminates the camel jockey slave problem in Qatar, it doesn't improve the economic and social conditions in places such as Sudan.

Computing technology often has important ethical impact. The young boys replaced by robotic camel jockeys are freed from slavery. Computing frees many of us from monotonous, boring jobs. It can make our lives better but can also make them worse. For example, we can conduct business online easily, but we're more vulnerable to identity theft. Machine ethics in this broad sense is close to what we've traditionally called *computer ethics*. In one sense of machine ethics, computers do our bidding as surrogate agents and impact ethical issues such as privacy, property, and power. However, the term is often used more restrictively. Frequently, what sparks debate is whether you can put ethics into a machine. Can a computer operate ethically because it's internally ethical in some way?

Implicit Ethical Agents

If you wish to put ethics into a machine, how would you do it? One way is to constrain the machine's actions to avoid unethical outcomes. You might satisfy

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

machine ethics in this sense by creating software that implicitly supports ethical behavior, rather than by writing code containing explicit ethical maxims. The machine acts ethically because its internal functions implicitly promote ethical behavior – or at least avoid unethical behavior. Ethical behavior is the machine's nature. It has, to a limited extent, virtues.

Computers are implicit ethical agents when the machine's construction addresses safety or critical reliability concerns. For example, automated teller machines and Web banking software are agents for banks and can perform many of the tasks of human tellers and sometimes more. Transactions involving money are ethically important. Machines must be carefully constructed to give out or transfer the correct amount of money every time a banking transaction occurs. A line of code telling the computer to be honest won't accomplish this.

Aristotle suggested that humans could obtain virtue by developing habits. But with machines, we can build in the behavior without the need for a learning curve. Of course, such machine virtues are task specific and rather limited. Computers don't have the practical wisdom that Aristotle thought we use when applying our virtues.

Another example of a machine that's an implicit ethical agent is an airplane's automatic pilot. If an airline promises the plane's passengers a destination, the plane must arrive at that destination on time and safely. These are ethical outcomes that engineers design into the automatic pilot. Other built-in devices warn humans or machines if an object is too close or the fuel supply is low. Or, consider pharmacy software that checks for and reports on drug interactions. Doctor and pharmacist *duties of care* (legal and ethical obligations) require that the drugs prescribed do more good than harm. Software with elaborate medication databases helps them perform those duties responsibly.

Machines' capability to be implicit ethical agents doesn't demonstrate their ability to be full-fledged ethical agents. Nevertheless, it illustrates an important sense of machine ethics. Indeed, some would argue that software engineers must routinely consider machine ethics in at least this implicit sense during software development.

Explicit Ethical Agents

Can ethics exist explicitly in a machine?[3] Can a machine represent ethical categories and perform analysis in the sense that a computer can represent and analyze inventory or tax information? Can a machine "do" ethics like a computer can play chess? Chess programs typically provide representations of the current board position, know which moves are legal, and can calculate a good next move. Can a machine represent ethics explicitly and then operate effectively on the basis of this knowledge? (For simplicity, I'm imaging the development of ethics in terms of traditional symbolic AI. However, I don't want to exclude the possibility that the machine's architecture is connectionist, with an explicit understanding of

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

the ethics emerging from that. Compare Wendell Wallach, Colin Allen, and Iva Smit's different senses of "bottom up" and "top down."[4])

Although clear examples of machines acting as explicit ethical agents are elusive, some current developments suggest interesting movements in that direction. Jeroen van den Hoven and Gert-Jan Lokhorst blended three kinds of advanced logic to serve as a bridge between ethics and a machine:

- *deontic* logic for statements of permission and obligation,
- *epistemic* logic for statements of beliefs and knowledge, and
- *action* logic for statements about actions.[5]

Together, these logics suggest that a formal apparatus exists that could describe ethical situations with sufficient precision to make ethical judgments by machine. For example, you could use a combination of these logics to state explicitly what action is allowed and what is forbidden in transferring personal information to protect privacy.[6] In a hospital, for example, you'd program a computer to let some personnel access some information and to calculate which actions what person should take and who should be informed about those actions.

Michael Anderson, Susan Anderson, and Chris Armen implement two ethical theories.[7] Their first model of an explicit ethical agent – Jeremy (named for Jeremy Bentham) – implements Hedonistic Act Utilitarianism. Jeremy estimates the likelihood of pleasure or displeasure for persons affected by a particular act. The second model is W.D. (named for William D. Ross). Ross's theory emphasizes prima facie duties as opposed to absolute duties. Ross considers no duty as absolute and gives no clear ranking of his various prima facie duties. So, it's unclear how to make ethical decisions under Ross's theory. Anderson, Anderson, and Armen's computer model overcomes this uncertainty. It uses a learning algorithm to adjust judgments of duty by taking into account both prima facie duties and past intuitions about similar or dissimilar cases involving those duties.

These examples are a good start toward creating explicit ethical agents, but more research is needed before a robust explicit ethical agent can exist in a machine. What would such an agent be like? Presumably, it would be able to make plausible ethical judgments and justify them. An explicit ethical agent that was autonomous in that it could handle real-life situations involving an unpredictable sequence of events would be most impressive.

James Gips suggested that the development of an ethical robot be a computing Grand Challenge.[8] Perhaps Darpa could establish an explicit-ethical-agent project analogous to its autonomous-vehicle project (www.darpa.mil/grandchallenge/index.asp). As military and civilian robots become increasingly autonomous, they'll probably need ethical capabilities. Given this likely increase in robots' autonomy, the development of a machine that's an explicit ethical agent seems a fitting subject for a Grand Challenge.

Machines that are explicit ethical agents might be the best ethical agents to have in situations such as disaster relief. In a major disaster, such as Hurricane

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

Katrina in New Orleans, humans often have difficulty tracking and processing information about who needs the most help and where they might find effective relief. Confronted with a complex problem requiring fast decisions, computers might be more competent than humans. (At least the question of a computer decision maker's competence is an empirical issue that might be decided in favor of the computer.) These decisions could be ethical in that they would determine who would live and who would die. Some might say that only humans should make such decisions, but if (and of course this is a big assumption) computer decision making could routinely save more lives in such situations than human decision making, we might have a good ethical basis for letting computers make the decisions.[9]

Full Ethical Agents

A full ethical agent can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will. Can a machine be a full ethical agent? It's here that the debate about machine ethics becomes most heated. Many believe a bright line exists between the senses of machine ethics discussed so far and a full ethical agent. For them, a machine can't cross this line. The bright line marks a crucial ontological difference between humans and whatever machines might be in the future.

The bright-line argument can take one or both of two forms. The first is to argue that only full ethical agents can be ethical agents. To argue this is to regard the other senses of machine ethics as not really ethics involving agents. However, although these other senses are weaker, they can be useful in identifying more limited ethical agents. To ignore the ethical component of ethical-impact agents, implicit ethical agents, and explicit ethical agents is to ignore an important aspect of machines. What might bother some is that the ethics of the lesser ethical agents is derived from their human developers. However, this doesn't mean that you can't evaluate machines as ethical agents. Chess programs receive their chess knowledge and abilities from humans. Still, we regard them as chess players. The fact that lesser ethical agents lack humans' consciousness, intentionality, and free will is a basis for arguing that they shouldn't have broad ethical responsibility. But it doesn't establish that they aren't ethical in ways that are assessable or that they shouldn't have limited roles in functions for which they're appropriate.

The other form of bright-line argument is to argue that no machine can become a full ethical agent – that is, no machine can have consciousness, intentionality, and free will. This is metaphysically contentious, but the simple rebuttal is that we can't say with certainty that future machines will lack these features. Even John Searle, a major critic of strong AI, doesn't argue that machines can't possess these features.[10] He only denies that computers, in their capacity as purely syntactic devices, can possess understanding. He doesn't claim that machines

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

can't have understanding, presumably including an understanding of ethics. Indeed, for Searle, a materialist, humans are a kind of machine, just not a purely syntactic computer.

Thus, both forms of the bright-line argument leave the possibility of machine ethics open. How much can be accomplished in machine ethics remains an empirical question.

We won't resolve the question of whether machines can become full ethical agents by philosophical argument or empirical research in the near future. We should therefore focus on developing limited explicit ethical agents. Although they would fall short of being full ethical agents, they could help prevent unethical outcomes.

I can offer at least three reasons why it's important to work on machine ethics in the sense of developing explicit ethical agents:

- Ethics is important. We want machines to treat us well.
- Because machines are becoming more sophisticated and make our lives more enjoyable, future machines will likely have increased control and autonomy to do this. More powerful machines need more powerful machine ethics.
- Programming or teaching a machine to act ethically will help us better understand ethics.

The importance of machine ethics is clear. But, realistically, how possible is it? I also offer three reasons why we can't be too optimistic about our ability to develop machines to be explicit ethical agents.

First, we have a limited understanding of what a proper ethical theory is. Not only do people disagree on the subject, but individuals can also have conflicting ethical intuitions and beliefs. Programming a computer to be ethical is much more difficult than programming a computer to play world-champion chess – an accomplishment that took 40 years. Chess is a simple domain with well-defined legal moves. Ethics operates in a complex domain with some ill-defined legal moves.

Second, we need to understand learning better than we do now. We've had significant successes in machine learning, but we're still far from having the child machine that Turing envisioned.

Third, inadequately understood ethical theory and learning algorithms might be easier problems to solve than computers' absence of common sense and world knowledge. The deepest problems in developing machine ethics will likely be epistemological as much as ethical. For example, you might program a machine with the classical imperative of physicians and Asimovian robots: First, do no harm. But this wouldn't be helpful unless the machine could understand what constitutes harm in the real world. This isn't to suggest that we shouldn't vigorously pursue machine ethics. On the contrary, given its nature, importance, and difficulty, we should dedicate much more effort to making progress in this domain.

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press

Acknowledgments

I'm indebted to many for helpful comments, particularly to Keith Miller and Vincent Wiegel.

References

1. H. Simon, "Re: Dartmouth Seminar 1956" (email to J. Berleur), Herbert A. Simon Col lection, Carnegie Mellon Univ. Archives, 20 Nov. 1999.
2. J. Lewis, "Robots of Arabia," *Wired*, vol. 13, no. 11, Nov. 2005, pp. 188–195; www. wired. com/wired/archive/13.11/camel.html?pg=1 & topic=camel&topic_set=.
3. J.H. Moor, "Is Ethics Computable?" *Metaphilosophy*, vol. 26, nos. 1–2, 1995, pp. 1–21.
4. W. Wallach, C. Allen, and I. Smit, "Machine Morality: Bottom-Up and Top-Down Approaches for Modeling Human Moral Faculties," *Machine Ethics*, M. Anderson, S.L. Anderson, and C. Armen, eds., AAAI Press, 2005, pp. 94–102.
5. J. van den Hoven and G.J. Lokhorst, "Deontic Logic and Computer-Supported Computer Ethics," *Cyberphilosophy: The Intersection of Computing and Philosophy*, J.H. Moor and T.W. Bynum, eds., Blackwell, 2002, pp. 280–289.
6. V. Wiegel, J. van den Hoven, and G.J. Lokhorst, "Privacy, Deontic Epistemic Action Logic and Software Agents," *Ethics of New Information Technology, Proc. 6th Int'l Conf. Computer Ethics: Philosophical Enquiry* (CEPE 05), Center for Telematics and Information Technology, Univ. of Twente, 2005, pp. 419–434.
7. M. Anderson, S.L. Anderson, and C. Armen, "Towards Machine Ethics: Implementing Two Action-Based Ethical Theories," *Machine Ethics*, M. Anderson, S.L. Anderson, and C. Armen, eds., AAAI Press, 2005, pp. 1–7.
8. J. Gips, "Creating Ethical Robots: A Grand Challenge," presented at the AAAI Fall 2005 Symposium on Machine Ethics; www.cs.bc. edu/~gips/EthicalRobotsGrandChallenge. pdf.
9. J.H. Moor, "Are There Decisions Computers Should Never Make?" *Nature and System*, vol. 1, no. 4, 1979, pp. 217–229.
10. J.R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–457.

https://doi.org/10.1017/CBO9780511978036.003 Published online by Cambridge University Press