Topoi (2013) 32:213–215
DOI 10.1007/s11245-013-9183-x

Introduction: Machine Ethics and the Ethics of Building Intelligent Machines

Marcello Guarini

Published online: 5 September 2013

1 A Brief History of Machine Ethics and Related Work

Machine ethics is an emerging field of research that studies the possibility of constructing machines that can mimic, simulate, generate, or instantiate ethical sensitivity, learning, reasoning, argument, or action. The machines in question may be physical or virtual; they may be stationary or mobile. Wallach and Allen (2009) provide an overview of the field, as do Anderson and Anderson (2007, 2011a, b). Robot nannies, robotic weapons systems, robots in elder care, virtual companions—these are just a sampling of some of the systems undergoing research and development. As the types of interaction humans and machines engage in become more and more complex, the concern that the design of such machines takes into consideration ethical norms (and maybe that the machines, in some sense, be guided by such norms) becomes pressing.

Speculations about the implications of increasingly intelligent machines have been around for some time. As Yampolskiy and Fox remind us in their contribution to this issue, as early as 1863 Samuel Butler considered the possibility that machines might be our successors. There have been various philosophical contributions not only to the discussion of intelligent machines generally, but also about the possibility of machines making use of or being guided by ethical considerations. Arguably the first monograph-length treatment of using computational techniques to construct models of moral reasoning is Danielson's (1992) *Artificial Morality: Virtuous Robots for Virtuous Games.* Much of the work in machine ethics before the current century could be seen as pioneering since there was little or no past work in the area to build on. That started to change near the turn of the century.

One of the marks of the development of a new field of research is a body of literature in which scholars cite each other and build on one another's work. Entire conferences or conference sessions devoted to a subject matter are other markers that a field devoted to that subject is developing. In 2002 the International Conference on Systems Research, Informatics and Cybernetics (in Baden-Baden, Germany) organized a workshop called Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. Conference organizer George Lasker invited Iva Smit to chair the original gathering, and Iva was joined by Wendell Wallach in several subsequent years as a co-chair for the workshop. Participants included Colin Allen, Bernard Stahl, Michael and Susan Anderson, Steve Torrance, Miranda Mowbray, Nick Bostrom, and others who have gone on to publish work in machine ethics. In 2005 in Arlington, Virginia, Michael and Susan Anderson organized the first AAAI symposium devoted to machine ethics. Authors in this area are now citing each other's work. A 2011 Cambridge Press anthology entitled *Machine Ethics*, edited by Michael and Susan Anderson, provides a useful snapshot of the range of work being done in this area.

Gianmarco Veruggio chaired the first international symposium on Roboethics in 2004 in Sanremo, Italy, and went on to organize a series of workshops in the area. Veruggio and Abney (2012) distinguish between roboethics and robot ethics. Roboethics is taken to be a branch of applied ethics that engages ethical issues arising from the building increasingly capable robots. Robot ethics refers to the programming or information structures that capture a robot's ethical code or sensibility. How significantly robot

M. Guarini (✉)
University of Windsor, Windsor, ON N9B 3P4, Canada
e-mail: mguarini@uwindsor.ca
URL: www.uwindsor.ca/guarini

ethics overlaps with machine ethics will depend in part on how broadly "robot" is defined. If the term "robot" is defined so broadly as to include virtual machines and stationary physical systems—which may sound strange to some ears—then there would be significant overlap. If we are more restrictive with the definition of robots and exclude virtual machines or stationary systems, then the subject matter of machine ethics ends up being broader than that of robot ethics. There is another dimension along which we might define overlap between these terms. Below I will discuss practical and reflexive approaches to machine ethics. If robot ethics is construed broadly enough that it encompasses not only the practical approach but the reflexive approach as well, then once again we have a significant overlap between the expressions.

It is difficult to predict which terms will catch on and what their future uses will be. Perhaps one of "machine ethics" or "robot ethics" will completely overtake the other. Perhaps both will stay in common use, with robot ethics being a species of the more general genus of machine ethics. Perhaps both expressions will be construed broadly and remain in common use as synonyms for one another. It may even be that both terms will fall out of use in favour of other terms. Yampolskiy and Fox (this issue) point out that there are many terms being used to refer to subject matters pertaining to ethics and intelligent machines. With anthologies by major presses entitled *Machine Ethics* (Anderson and Anderson 2011c) and *Robot Ethics* (Lin et al. 2012), it is likely that these terms will be in common use for at least the near future.

2 Approaches

There are different motivations for doing machine ethics. As I have done elsewhere (Guarini 2011), I will refer to these as *practical* and *reflexive*. Those who want to build machines for some application are practically motivated; those who want to build machines or reflect on how they could be built to better understand what ethics is, or what it could be, are reflexively motivated. Of course, the motivations are not mutually exclusive: one could certainly be motivated in both ways. We can also speak of practical and reflexive *approaches* to machine ethics. Both approaches have philosophical dimensions.

The practical approach is not just about traditional engineering. Actually figuring out how to construct the physical or virtual machine that is ethically constrained or reasons about ethics or the like will require some thought about what sorts of ethical constraints (or reasoning or sensitivity…) are appropriate. It will also require figuring out how to specify the problems and procedures for arriving at solutions in enough detail that they can be implemented. Philosophers can make contributions to both of these challenges. So while there is much in the practical approach that is empirical and computational in nature, it does not preclude philosophical contributions.

The reflexive approach more obviously involves philosophy—reflecting on the nature of ethical sensitivity, awareness, reasoning, or argument is widely conceived as philosophical in nature. That said, when one takes up the design stance (Dennett 1978, chapter 1) for the purpose of trying to better understand ethics, the project is almost unavoidably informed by empirical and computational considerations.

I have heard some refer to the "engineering" side of machine ethics and the "philosophical" side. There is something to that, but I think reference to practical and reflexive approaches better captures what is going on in this area of research. The practical approach recognizes that the empirical, computational, and engineering work is accompanied by philosophical reflection, and the reflexive approach understands that philosophical reflection can be informed by empirical, computational, and engineering considerations.

Of course, there are ethical considerations about the very building of increasingly intelligent machines in general, and about building ethical machines that may be able to reason about ethical matters. These questions are not generally seen as part of machine ethics; they are concerned with the ethics of building intelligent machines, something Susan Anderson (2011) sees as part of machine metaethics. There are important questions to be asked about how intelligent we should make machines, and which ones (if any) should be built. While this is not the same thing as taking a practical or reflexive approach to machine ethics, there are interesting interconnections. If we make machines that have the capacity to build other machines (physical or virtual), the question arises as to what kinds of machines, if any, we should allow machines to make. Ethical questions about the building of intelligent machines may then inform both the practical and reflexive approaches to machine ethics (and *vice versa*).

3 The Contributions

The lead-off paper is by Roman Yampolskiy and Joshua Fox. They are concerned with machines that may be more intelligent than human beings, which machines of that sort should be built, and the need for safety engineering to prevent such machines from being able to do harm. Among other things, the paper is a contribution to the ethics of building super intelligent machines. Yampolskiy and Fox argue that the intelligent machines they consider should be treated neither as moral agents nor as moral patients.

In his paper, Thomas Powers explores what it is that might make a computer an ethical agent. The paper also examines what it is to be an ethical patient. While many thinkers recognize that there might be beings who are ethical patients but not agents (say, very young children), it is generally thought that all agents also qualify as patients. Powers raises the possibility that some computers may qualify as ethical agents but fail to qualify as patients (even if traditional ethical theories do not recognize this possibility). This is a reflexive exploration of machine ethics.

Gregory Reed and Nicholas Jones take a practically motivated approach to machine ethics in their contribution. They are interested in whether it is possible to construct a metric of evil, one which might aid in military decision making. Both engineering and philosophical considerations are readily apparent.

Paul Bello and Selmer Bringsjord are both practically and reflexively motivated. They argue that the study of how humans attribute causal responsibility could help us to build into machines an understanding of such abilities so that machines are better able to interact with us. There is a real (practical) concern here with how we might go about building ethically responsive machines; there is also a reflexive concern with how attempting to computationally model human attributions of causal responsibility might lead to a more psychologically realistic understanding of human ethical rationality.

My own contribution is reflexively motivated. Tools from computational neural modeling are used both to reflect on new ways of thinking about ethical similarity between cases, and to re-imagine debates between particularists and generalists in their discussion of the nature of moral sensitivity, reasoning, or argument.

Finally, it is not by happenstance that this issue appears now: the year 2012 was the centenary of Alan Turing's birth, marked by various conferences around the world, and his seminal paper, "Computing Machinery and Intelligence" (1950), is (un)timely reviewed by Cristiano Castelfranchi in this very same issue. Turing's work influenced both computer science and philosophy. While machine ethics as an emerging field did not exist in his time, his vision of bringing together computational and philosophical reflection fostered the development of an intellectual environment that helped to make so much possible, including machine ethics.

Acknowledgments All papers in this volume are refereed, and I thank all those who helped with the refereeing. I also thank Wendell Wallach for his discussions about the early days of machine ethics. Some of the history provided above was informed by my conversations with Wendell. I take responsibility for omissions or any other shortcomings of that brief history. Thanks also go to Fabio Paglieri for encouraging and supporting the development of this collection.

References

Anderson S (2011) Machine metaethics. In: Anderson M, Anderson S (eds) Machine ethics. Cambridge University Press, pp 21–27

Anderson M, Anderson S (2007) Machine ethics: creating an ethically intelligent agent. AI Mag 28(4):15–26

Anderson M, Anderson S (eds) (2011a) General introduction. Machine ethics. Cambridge University Press, pp 1–4

Anderson M, Anderson S (eds) (2011b) Introduction. Machine ethics. Cambridge University Press, pp 7–12

Anderson M, Anderson S (eds) (2011c) Machine ethics. Cambridge University Press, Cambridge

Danielson P (1992) Artificial morality: virtuous robots for virtuous games. Routledge, New York

Dennett D (1978) Brainstorms: philosophical essays on mind and psychology. MIT press, a Bradford Book, Cambridge, MA

Guarini M (2011) Computational neural modeling and the philosophy of ethics: reflections on the particularism-generalism debate. In: Anderson M, Anderson S (eds) Machine ethics. Cambridge University Press, Cambridge, pp 316–334

Lin P, Abney K, Bekey GA (eds) (2012) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, MA

Turing A (1950) Computing machinery and intelligence. Mind 59(236):433–460

Veruggio G, Abney K (2012) Roboethics: the applied ethics for a new science. In: Lin P, Abney K, Bekey GA (eds) Robot Ethics. MIT Press, Cambridge, MA, pp 347–363

Wallach W, Allen C (2009) Moral machines: Teaching robots right from wrong. Oxford University Press