Human Systems Management 29 (2010) 65–68
DOI 10.3233/HSM-2010-0712
IOS Press

Book review

Philosophy *plugged*: How robotics informs ethics*

Alan E. Singer
E-mail: singerae@appstate.edu

1. Introduction

Designers of artificial moral agents (AMA's) or ethical (ro-)*bots* will be informed by this book. However, it will also challenge moral philosophers and anyone involved in teaching ethics. Indeed, an alternative subtitle: "teaching *ethicists* right from wrong" would be quite appropriate. The book demonstrates quite convincingly that "you don't really know how something works if you can't build it", so that "robotocists are doing philosophy, whether or not they think this is so" [5]. Yet this "philosophy" is *plugged:* an experimental and constructive "computational philosophy" that fits well with the notion of knowledge as coordination-of-action (e.g. [12]) and the associated position that the physical and mental worlds (are becoming) one and the same[1] In addition, the task of AMA design and construction repeatedly spins-off sharply-framed questions that are both philosophical and technological.

When Ethics is plugged-in, it looks and feels quite different from the penned works of Kant, Mill, Bentham, or the Bible. In part this is because AMA development it is to a large extent a project of the military-indutrial complex, being "done" outside the public gaze (cf. [11]). For example, a military project is discussed in the book that involves installing (instilling) a "functional morality" into a robot machine gun. The guns have to be re-programmed with ethics so they stop killing friendlies or "innocent" civilians, but concentrate firepower on the bad guys. This is one example of the general principle that, in the evolution of AMA's, autonomy precedes "sensitivity", just as consciousness arguably precedes conscience in evolution.

The idea of autonomy in robots [6] points to a "hard takeoff", predicted at *c.* 2020–2050, after which *self*-programming AMA's will be able to make themselves ever more intelligent (and conceivably more moral) than humans. Such robots would also become capable of re-programming each other, including those guns. On a more sociable note, humans might be able to legally marry robots, or contract for "everything but marriage". After all, there is a well-considered view (cf. Savelescu quoted in [10]) that "if other beings possess rationality and the ability to cooperate and to empathise . . . then we should treat them no differently than other human beings". Equally controversial is the possible co-production of super-good or super-bad "enhanced" humans, whose brain cells are fused to AMA-silicon chips (wet AI).

2. Engineering informs ethics

Meanwhile, ethics-*unplugged* is in quite serious trouble. Almost all applied ethics textbooks kick off with an overview of "grand theories" such as egoism, utilitarianism and deontology. However, Wallach and Allen note that "it is not possible to see a clear way to implement an ethical theory as a computer-program" so "one might wonder whether these (theories) play a guiding role for human action". Mindful that one can't replace something with nothing, several alternative ideas about

*Review of "*Moral Machines: Teaching Robots Right From Wrong*" by Wendell Wallach & Colin Allen, Oxford University Press, 2009.

[1]Baruch Spinoza considered "the physical and mental worlds one and the same" and that there is a "universal substance consists of both body and mind" (*Wikipedia*/Spinoza). Milan Zeleny has pointed out that "information" is becoming "in-*formation*" (i.e. physical production).

how ethics might work are explored. These include notions of bottom-up, evolutionary and incremental ethics, where engineering seems to inform philosophy, as follows.

2.1. Bottom up ethics

Genghis is an insect-like robot. It does not have much of a brain, but it certainly appears to know what it's doing. Each leg takes its cue from the other legs, with a few local features. Genghis' "knowledge" is thus fully expressed as coordination-of-action, as there is not much else. Yet, Genghis moves around better than its more brainy competitor *bots*. (Although it's not mentioned in the book, this discovery *c*. 1990 seems to have inspired management consultants to preach "subsumptive organizational architecture", meaning that department heads should talk directly with each other, with no HQ.) In some ways, practical ethics and moral judgment do seem to be more like Ghengis than Kant (or God). Socially adept responses and authentic social skills do constitute an important part of ethics in business, as elsewhere. Perhaps, therefore, applied ethicists should pay a bit more attention to these aspects of behavior.

2.2. Ethical incrementalism

The Learning Intelligent Distribution Agent (LIDA) is an Autonomous General Intelligent (AGI) system built by the US Navy to make human resource related decisions including deployments. Here Ethical decision making (EDM) becomes a series of selections of (internal and external) micro-actions, rather than choices amongst alternative major projects (this is akin to the "logical incrementalism" described by business and policy consultants since *c*.1980). Inside LIDA, lines of software known as *codelets* scan a virtual workspace in which all inputs are represented [2]. The codelets (which as the book notes, are quite similar to the demons in a 1970's cognitive model called Pandemonium, but also the "agents" in Minsky's [9] *Society of Mind*) scan this workspace for information that should be brought to the attention of the wider system, or brain in a competition for attention that lasts about 0.1 second, a "winning" piece of information emerges. This "winner" is then "broadcast" throughout the system. The next step within each cycle is to either (i) act, or (ii) reflect more, or (iii) add something to a mental model that is always under construction in semantic memory. So, even though LIDA does not execute programs of TD moral reasoning, it detects "morally-relevant inputs" and acts on them. This notion also seems to merit more attention from philosophers and students of applied ethics.

2.3. Evolutionary ethics

"Survival" is held to be a basic value for autonomous systems (and business enterprises). Expanding on this (p. 193), a set of "easy basic values" are identified in the book (cf. [7]). These are: (a) keep yourself healthy, (b) preserve patterns that have been valuable, (c) create diversity. These should be relatively easy to program into an AMA/AGI. They are also good for humans, although the last two differ from classical Platonic human goods. The "Hard values" in contrast include preserving other life (at least friendly life?) and making others happy. These are not so obvious, nor natural, and they normally have to be taught by society or imposed by an authority (e.g. "Thou shall not kill").

In chapter 8 (merging top-down and bottom-up) it is pointed out that the extent to which a robot obeys moral laws (such as Asimov's laws of Robotics [1]) might be used as a fitness criterion in an "evolutionary competition". As in Axelrod's high-impact "evolution of cooperation" work in the 1980's, winners could be selected for breeding in the next round. Such experiments would then put flesh (or nuts) on the concept of "survival of the most moral" and would again help humans to understand the effect of specific moral laws on their own survival.

3. Ethics informs engineering

Despite the observation that traditional moral philosophy appears to distract engineers and programmers, some core philosophical themes such as consequences, logic and rationality, virtue and emotion all appear to have guided (or anchored) the AMA project, as follows.

3.1. Consequentialism

Since computer simulations enable better identification and forecasting of moral effects or consequences (e.g. traffic delays, pollution, etc.) it should be possible to build a powerful consequentialist AMA. This might be egoist (self-interested, perhaps with the "easy" values) or utilitarian (weighing the interests of all stakeholders). In either case, such an agent would be capable of a more informed moral judgment than an unaided

human egoist or utilitarian, respectively. That is, it would be able to pass a moral Turing Test.

3.2. Deontology

The book also notes (p. 95) that "a very powerful computer might be able to determine whether its current goal would be blocked if all other agents were to operate with the same motive or *maxim*". That is, it could execute a version of the Golden Rule (a kind of Kant-*plugged*). This again holds out the prospect of super-moral AMAs, because humans have to be cajoled into, or extrinsically rewarded for, following such rules. On the other hand, such an AMA would immediately and permanently shut down that *bot*-gun. Another Kant-inspired "super-moral" line of research (by a Dutch group) involves the use of theorem-proving software to assess the adequacy of a block of software code for creating its intended outcome.

3.3. Virtue

The engineering mind has also been steered towards classical virtue ethics. Plato pointed to secular virtues such as wisdom, courage, moderation and justice. A future AMA might be able to emulate some aspects of these. Aristotle drew a distinction between intellectual virtues that can be taught (or programmed in a top-down sense), such as loyalty and moderation vs. the moral virtues such as humor and politeness that are typically acquired through practice and habit. This classical distinction remains useful today. Indeed there is considerable discussion in the book of the linkages between, on the one hand, neural nets, connectionist psychology and particularist ethics (e.g. [4]) and on the other hand, the top-down rules and guides found in the grand theories and ethical principlism (e.g. [3]).

3.4. Emotion

As with consequences, logics and virtues, many aspects of emotion (or emotional intelligence) are also programmable or capable of being learned by an AMA. These include (i) the ability to detect and respond intelligently and expressively to others' facial expressions or body posture, and (ii) interpreting other's intentions in context, or responding sympathetically and appropriately to others' predicaments. A tougher challenge involves making a robot behave as if it was experiencing or anticipating its own *quasi*-emotions. For example, an emotionally-intelligent robot gun might avoid friendly fire if it anticipates the pain this would cause to *itself*. This "pain" would only have to be some internal state with "valence" (a $\pm$ parameter), such as opposing the robot's "easy" values, or interfering with its goal-attainment, or slowing it down. Metzinger's [8] Phenomenal-Self model is mentioned in this context (p. 205). Here, a robot or human is able to "see" its own somatic responses. If peripheral components somehow responded directly to emotionally (and morally) relevant inputs, it might be able to compute *quasi*-emotions and adjust its behavior accordingly. (This is like the autonomous nervous system in animals and humans, but also fits well with LIDA and Genghis.)

4. Policy

There are several discussions in the book of the macro-ethical (policy) issues involved in living with robots, or *as* robots, in the future. In line with other accounts of the social effects of technologies (e.g. nanotechnology, the precautionary principle, etc.) the potential for both good or harm is explored. On the optimistic side, there is a reference to an *invisible hand of system interactions,* the idea that the operation of many self-sustaining easy-value holding AGIs/AMA's will somehow lead to overall (macro-) good even if they lack values such as helping others. In contemplating this update of Adam Smith, however, it should not be forgotten that Smith emphasized the role of human moral communities. On the side of harm, we are warned of a possible "social tsunami".

5. Conclusion

Perhaps the biggest difference between 1776 and 2010+ lies not in the invisible hand, but in the thorough blurring of boundaries and categories. As wet-AI advances, even categories such as robotics, nanotechnology and ecology, not to mention ethics, economics and technology will become increasingly reformed. Physical devices such as neuro-prosthetics and nano-tech implants place symbolic mind-like systems inside the body (i.e. Human Enhancement) whilst the classical "mightier" pen is already fully inside the sword, or the gun. Perhaps, then, the most profound way in which technology informs philosophy is by explicating the idea that the physical and mental worlds are ultimately one and the same (*neutral monism*). They are undoubtedly *becoming* that way.

Wallach and Allen also point out that aircraft and birds fly in different ways. Accordingly, when an AMA is built, even a perfectly moral (or deadly) one, we still might not know how human morality really works. According to the "robust view of ethics", for example, a real moral agent (such as a human) must have "real" feelings or sentience. It must experience *qualia,* the qualitative aspect of emotion. According to theological ethics, human virtues include religious faith, with a belief in the soul (of humans, at least). A robot can of course express these beliefs and act *as if* it holds them; it might also be sensitive to others who hold these beliefs; but can it actually *have* faith and a soul? This question remains important to many, even though trying to answer it may be a "Monkish pursuit" (p. 215). Perhaps, then, the most enduring way in which robotics has informed ethics is by making ideas such as "faith" and "soul" appear more sharply defined and well-focussed than ever before. It follows that if "building AMA's highlights the need for a richer understanding of human morality" it also helps to fulfill that same need.

References

[1] I. Asimov, *I Robot*, Gnome Press, NY, 1950.

[2] B. Baars, *In the Theatre of Consciousness: The Workspace of the Mind*, OUP, Oxford, 1997.

[3] T.L. Beauchamp and J.F. Childress, *Principles of Biomedical Ethics*, 4th edn, OUP, NY, 1994.

[4] J. Dancy, Can a particularist learn the difference between right and wrong? *20th World Congress of Philosophy*, Boston, 1998.

[5] D.C. Dennett, Cog as a thought experiment, *Robotics & Autonomous Systems* **20**(2–4) (1997), 251–256.

[6] L. Floridi and J. Sanders, On the morality of artificial agents, *Minds & Machines* **14**(3) (2004), 349–379.

[7] B. Goertzel, Thoughts on AI morality, *Dynamic Psychology*. www.goertzel.org, 2002.

[8] T. Metzinger, *Being No-One: The Self-Model Theory of Subjectivity*, MIT Press, Cambridge, 2004.

[9] M. Minsky, *The Society of Mind*, Simon and Schuster, NY, 1988.

[10] P. Snow, Woe, Superman? *Oxford Today* **22**(1) (Michaelmas) (2009), 13–15.

[11] A. Toffler and H. Toffler, War and Anti-war, *Survival At the Dawn of the 21st Century*, Little Brown, NY, 1990.

[12] M. Zeleny, *Human Systems Management*, World Scientific, London, 2007.