The Ethics of AI

Introduction

There are two questions about the ethics of **artificial intelligence** (AI) which are central:

1. How can we build an ethical AI?
2. Can we build an AI ethically?

The first question concerns the kinds of AI we might achieve — moral, immoral or amoral. The second concerns the ethics of our achieving such an AI. They are more closely related than a first glance might reveal. For much of technology, the National Rifle Association's neutrality argument might conceivably apply: "guns don't kill people, people kill people." But if we build a genuine, autonomous AI, we arguably will have to have built an artificial **moral agent**, an agent capable of both ethical and unethical behavior. The possibility of one of our artifacts behaving unethically raises moral problems for their development that no other technology can.

Both questions presume a positive answer to a prior question: Can we build an AI at all? We shall begin our review there.

The Possibility of AI

Artificial intelligence as a research area arose simultaneously with the electronic computers (Turing, 1948). AI aims at producing an intelligent machine by the construction of an appropriate computer program; the assertion of the possibility of this is known as the **strong AI thesis**. Turing (1950) proposed replacing the question whether a machine could be intelligent, by another: is it possible to program a machine so that its verbal behavior would be indistinguishable from human verbal behavior? This has become known as the **Turing Test** for intelligence. Turing thought his test would be passed by the year 2000. The continued failure to do so has paralleled continued debate over the possibility of doing so and also over the adequacy of the test.

Joseph (Weizenbaum, 1966) produced a natural language understanding program, ELIZA. This program had a small set of canned phrases and the ability to invert statements and return them as questions. For example, if you type "I am unhappy," it could respond "Are you unhappy often?" The program, however, is quite simple and, on Weizenbaum's own account, stupid. Nevertheless, Weizenbaum (1976) reported that the program's behavior was sufficiently human-like that it confused his secretary for some time; and it encouraged others to convert it into a kind of virtual psychologist, called DOCTOR, leading some to prophesy the arrival of automated therapy. Weizenbaum responded to these events with despair, swearing off any further AI research and declaring the profession unethical (more of which below).

Around this time Hubert Dreyfus launched an attack upon the possibility of an AI passing the Turing Test (Dreyfus, 1965). His arguments emphasized the many qualitative differences between

human thought and computation, including our embodiment (versus program portability), our intuitive problem-solving (versus rule-following), and the sensitivity of our judgements to mood (versus cold calculation). If these arguments were right, our computers could never achieve intelligence. However, Dreyfus (1994) ended up conceding that **artificial neural networks** (ANNs) potentially overcome these objections. Since ANNs are provably equivalent to ordinary computers (assuming they cannot overcome known physical constraints to perform infinite-precision arithmetic; see Franklin and Garzon, 1991), this indirectly conceded the possibility of an AI. (Korb, 1996, presents this argument in detail.)

Whatever the difficulties in tackling the Turing Test, we can legitimately wonder whether even passing it would suffice for intelligence. The best known argument against the adequacy of the Turing Test was launched by John Searle (1980), in the **Chinese Room Argument**. Searle began by granting the possibility of passing the Turing Test. Suppose we understand human natural language processing so well that we can precisely mimic it in a computer program. In particular, imagine a program able to understand and generate Chinese to this level. Searle chooses Chinese because *Searle* doesn't understand it. Write that program on paper; or rather, rewrite it in English pseudo-code so that Searle can understand it. Put the program, Searle, paper and ink in a giant room with two slots, one for input and one for output. If a Chinese speaker writes a squiggle on paper and inputs it, Searle will simulate the program, and after much to-ing and fro-ing, write some squoggle and output it. By assumption, Searle is participating in a Chinese conversation, but, of course, he doesn't understand it. Indeed, Searle's point is that nothing whatever in the Chinese Room *does* understand Chinese: not the Searle, not the paper with pseudo-code printed on it, nothing. Therefore, Searle concludes, there is no Chinese understanding going on and so passing the Turing Test is logically inadequate for intelligence.

The most popular response amongst AI researchers is to insist that it is no one thing within the room that is responsible for intelligence, rather it is the system (room) as a whole. Many systems have properties that emerge from the organization of their parts without inhering in any subpart, after all. All living organisms are examples of that. So why not intelligence? Harnad (1989), and many others, have responded by pointing out that intelligence requires semantics and the Chinese Room cannot have any successful referential semantics. For example, if the Chinese interlocutor were asking the Room about her fine new shirt, the Room would hardly have anything pertinent to say. For a program to display *human-like* intelligence it must be embodied in a robot with human-like sensors and effectors. Searle, on the other hand, thinks that intelligence and consciousness are necessary for each other (Searle, 1992). Functionalists would agree, although for different reasons. **Functionalism** asserts that the mind, including conscious states, depend only upon the biological functions implemented by the brain, including information-processing functions. Any system, wet or silicon, which implements those functions will, therefore, necessarily have a mind and consciousness (Dennett, 1991). This amounts to the view that strong AI, while strictly speaking false, can be largely salvaged by requiring that our computer programs be supplemented by bodies that support human-like behavior and semantics. The result will be a conscious, intelligent artifact, eventually. Assuming this to be so, let us reconsider the ethics of the matter.

Is AI Ethical?

Weizenbaum claimed that AI research is unethical. His reasons were not simply his personal despair at finding stupid AI programs pronounced smart. His argument (crudely put) was one which has repeatedly found favor in Hollywood: that once we build a genuine AI it will necessarily be intelligent and autonomous; that these AIs will lack human motivations and be incomprehensible

to us as well, as any large computer program must be; in other words, these AIs will be out of control and dangerous. The danger in science fiction is frequently manifested in a war between robots and their would-be masters.

It may be difficult to take Hollywood and its arguments seriously. But the potential dangers of an uncontrolled AI can be, and have been, put more sharply (Bostrom, 2002). The strong AI thesis, in effect, claims that if we were to enumerate all possible Turing machines from simpler to more complex, we would find machines which are isomorphic to you and me somewhere early in the list, one isomorphic to Einstein a little farther out, and perhaps the yet-to-be-encountered Andromedans quite a lot farther out. But there is no end to the list of Turing machines and no end to their complexity. Humans have various corporeal restrictions to their potential intelligence: their brains must fit through the birth canal, subsequent maturation can last only so long, etc. Although incorporated AIs will also face some restrictions, such as the speed of light, these are not nearly so severe. In short, once the first AI is built, there is no obvious limit to what further degrees of intelligence can be built. Indeed, once the first AI is built it can be replicated a great number of times and put to the problem of improving itself. Each improvement can be applied immediately to each existing robot, with the likely result that improvements will come thick and fast, and then thicker and faster, and so on. In what has been dubbed the **technological singularity**, we can expect that roughly as soon as there is a legitimate AI, there shall also be a **SuperIntelligence** (SI) (Good, 1965,Vinge, 1993, Bostrom, 1998; see also **cyberpunk** literature such as Gibson, 1984). An uncontrollable SI would be a very serious threat indeed. If such is the prospect, there can be little doubt that AI research is unethical.

A generic counterargument applies to any technology. Richard Stallman has famously argued that software will be free (a paraphrase of Stallman, 1992). Whether or not that is so, it seems clear that knowledge “will” be free: once a scientific research program appears feasible, it is already too late to stop it. If one party refuses to proceed in developing a technology, be it nuclear weaponry, therapeutic cloning or AI, that will simply leave the way open for others to get there first. Unless your motives are unethical, it cannot be unethical actually to get there first, barring some verifiable agreement by all parties to restrain themselves. So, it behooves us to consider whether a controllable AI might be possible.

Controlling AI

Isaac Asimov got there first. Asimov wrote a lengthy series of robot stories in which a unifying theme was his “Three Laws of Robotics” (Asimov, 1950):

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given to it by human beings except where such orders would conflict with the First Law.

3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Unfortunately, many of his plots revolved around the ambiguities and conflicts arising between these laws (even leading to the introduction of a “zeroeth” law, that a robot may not injure humanity as a whole, etc.). Guaranteeing that incorporating such laws in the psychological foundations of our robots would not give rise to problems, and potentially to loss of control, would require a semantic sophistication that is currently beyond human capacity.

The failure of such laws to maintain control is one kind of difficulty. But there is another. If you reread the laws in the language of some fascists, substituting "subhuman" for "robot", this will immediately become apparent: Asimov's laws have nothing to do with promoting ethical behavior; they are all about the selfish protection of human interests. If we were talking about the development of any neutral technology, which would be used for good or ill depending solely upon the motivations of its users, then this narrow focus would be natural. But, as we have seen, it is at least arguable that if AIs are achievable at all, they will be autonomous, with an independent set of motivations, and perhaps consciousness. They will more likely be artificial moral agents, than neutral slabs of technology. If we wish them to respect our rights, we will likely have to respect theirs first.

Future Trends

If we take the possibility of creating an artificial moral agent seriously, a possible resolution of the ethical problems readily suggests itself: we can build artficial agents which are capable of moral behavior and which *choose* to act ethically. The possibility of constructing moral agents has received attention recently (e.g., Johnson, 2007; Floridi and Sanders, 2004).

Allen et al. (2000) raise the question: when could we know we have created an artificial moral agent? They propose using a **moral Turing Test**. When one of our creations has passed it, we will have created an artifact which is morally indistinguishable from some (normal) humans. Human history suggests that this test may be passed too easily to be of interest to us. In particular, if we are creating a being whose intellect considerably exceeds our own, we are unlikely to be satisfied unless its ethics also considerably exceeds our own.

Of course, how to achieve such a goal depends upon what the right account of ethics may be. There are three leading types of normative ethics: deontic systems with rules of behavior (such as Moses' laws or, were they designed to be ethical, Asimov's laws); virtue ethics, which identifies certain moral characteristics (e.g., honor, integrity) which moral behavior should exemplify; consequentialism (including **utilitarianism**; Smart, 1973), which identifies moral value not from intrinsic properties of the action, but from its consequences. The debate between these views has been raging for more than two thousand years and is unlikely to be resolved now. A practical response for an artificial ethics project is to consider which of these is amenable to implementation. The difficulties with Asimov's laws show us that implementing any deontic ethics requires us to first solve our problems with natural language understanding, which is effectively the same as solving our problems with designing an AI in general. But our main problem here is how to build ethics *into* our AI, and this must be solved before we have created that AI. Similar difficulties apply to virtue ethics.

In the last few decades there has been considerable improvement in automated decision analysis using **Bayesian networks**, finding many hundreds of useful applications (Howard and Matheson, 1984, Jensen, 2001). These networks provide relatively efficient means of automating decision making so as to maximize expected utility in an uncertain world, which is one leading theory of what it means to act rationally (Russell and Norvig, 2003). This technology, or rather some future extension of it, promises to enable autonomous robots implementing arbitrary utility structures (motivations, goals), without the necessity of resolving all possible ambiguities or conflicts we might find in rules of any natural language.

Thus, for example, we might enforce a non-linguistic correlate of Asimov's laws upon such robots. However, if the robots are indeed autonomous seats of moral agency, this could be no more ethical than imposing such rules of enslavement upon any subpopulation of humans. A more promising approach is to build the robots so that they are ethical. As agents, they must

have some utility structure. But it needn't be one which is solely concerned with maximizing their private utility (implementing **egoism**); instead, it could be utilitarian, maximizing expected utilities across the class of all moral agents, in which case the well-being of humans, separately and collectively, would be one of their concerns.

There are many difficulties in the way of a project to implement an artifical utilitarian agent. Allen et al. (2000) argue that this artificial morality project requires computing all the expected consequences (and so utilities) of actions and that this is intractable, since there is no temporal or spatial limit to such consequences; further, any horizon imposed on the calculation would have to be arbitrary. But this objection ignores that utilitarianism advocates maximizing *expected* utility, not *absolute* utility. No reasonable ethics can demand actions (or calculations) beyond our abilities; what we expect to arise from our actions is always limited by our abilities to formulate expectations. And those limits fix a horizon on our expectations which is the opposite of arbitrary. Nevertheless, the history of ethics suggests that the most intransigent difficulties in the way of the project will be the theoretical debates around its value, rather than the practical problem of developing and applying the technology.

Conclusion

With all of our technologies there are serious moral issues about their value, and especially about the value of the uses to which we put them. If those uses are likely to be unethical, then the ethics of those developing them can be put into doubt, at least if there is any alternative. For AI matters are even worse: since AIs will be autonomous actors, and since once they arise they will rapidly exceed our abilities, they may put *themselves* to unethical uses. However, there is a real option of designing them to be ethical actors in the first place, as well as a real technology to support such an effort. If realized, then our robotic offspring, as well as their future descendants, need not be feared. Thus, we might find our robotic grandchildren caring for their senescent grandparents one day, without either one dominating the other.

References

Allen, C., G. Varner, and J. Zinser (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical AI 12*, 251–261.

Asimov, I. (1950). *I, Robot*. Gnome Press.

Bostrom, N. (1998). How long before superintelligence? *International Journal of Futures Studies 2*.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology 9*. http://jetpress.org/.

Dennett, D. (1991). *Consciousness Explained*. Boston.: Little, Brown and Company.

Dreyfus, H. (1965). Alechemy and artificial intelligence. Technical Report P. 3244, RAND Corporation.

Dreyfus, H. (1994). *What Computers Still Can't Do* (third ed.). MIT Press.

Floridi, L. and J. W. Sanders (2004). On the morality of artificial agents. *Minds and Machines 14*, 349–379.

Franklin, S. and M. Garzon (1991). Neural computability. In O. Omidvar (Ed.), *Progress in Neural Networks*. Norwood, NJ: Ablex Publishing Corp.

Gibson, W. (1984). *Neuromancer*. Ace Books.

Good, I. (1965). Speculations concerning the first ultraintelligent machine. In *Advances in Computers*, Volume 6, pp. 31–88. Academic Press.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence 1*, 5–25.

Howard, R. and J. Matheson (1984). Influence diagrams. In R. Howard and J. Matheson (Eds.), *Readings on the Principles and Applications of Decision Analysis*. Menlo Park, CA: Strategic Decisions Group.

Jensen, F. (2001). *Bayesian networks and decision graphs*. New York: Springer.

Johnson, D. G. (2007). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*.

Korb, K. B. (1996). Symbolicism and connectionism: Ai back at a join point. In D. L. Dowe, K. B. Korb, and J. J. Oliver (Eds.), *Information, Statistics and Induction in Science*, pp. 247–257. World Scientific.

Russell, S. and P. Norvig (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*, 417–457.

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, Mass.: MIT Press.

Smart, J. J. C. (1973). An outline of a system of utilitarian ethics. In *Utilitarianism: For and Against*. Cambridge University Press.

Stallman, R. (1992/2002). Why software should be free. In *Free Software, Free Society: Selected Essays of Richard M. Stallman*. Free Software Foundation.

Turing, A. (1950). Computing machinery and intelligence. *Mind 59*, 433–460.

Vinge, V. (1993). The coming technological singularity. *Whole Earth Review*.

Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between men and machines. *Communications of the ACM 9*, 36–45.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. New York: W.H. Freeman.

Key Terms

Artificial intelligence.

1. The research field which investigates methods of improving the apparent intelligence of computer programs, including such subfields as: planning; visual processing; pattern recognition; natural language understanding; machine learning.

2. A computer system or robot which has achieved human-level intelligence (or greater intelligence), displayed across some wide range of behaviors.

Artificial neural networks (ANNs). A computational method based upon a simple model of biological neural processing.

Bayesian networks. A technology for automating probabilistic reasoning, commonly augmented with decision nodes to support decision making.

Chinese Room Argument. John Searle's argument that the Turing Test fails to establish intelligence, since it fails to establish any semantic understanding.

Cyberpunk (from "cybernetics" and "punk") is a type of science fiction which emphasizes the possibility of future disasters centered around technological change, especially future information technologies.

Egoism. The ethical view which holds that one ought to do what is in one's own self interest.

Functionalism. The thesis that mental states (including conscious states) are identifiable strictly in terms of the functional roles which they play, including information-processing roles. It follows that there may be many possible ways to realize these functions; in particular, both biological and silicon realizations may be possible.

Moral agent. An agent which is capable of moral behavior, implying the ability to behave both morally and immorally.

Moral Turing Test. A test of an artificial agent's ability to behave in ethically demanding tasks in a way indistinguishable from some (normal) humans.

The Strong AI thesis is the claim that designing and implementing some computer program could suffice for creating an artificial intelligence.

SuperIntelligence. A computer system or robot which has achieved greater than human-level intelligence, displayed across some wide range of behaviors.

Technological singularity. A point in time when technological change accelerates as if at a point of singularity (diverging rapidly to infinity). As applied to artificial intelligence, this has been called the "intelligence explosion."

Turing Test. The Turing Test proposes the indistinguishability of computer verbal behavior from human verbal behavior as a criterion of intelligence.

Utilitarianism. The ethical view which holds that one ought to do what is in the global interest of some class of agents.