Ethics of AI

Kevin B. Korb
Monash University, Australia

INTRODUCTION

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

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 first 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. Alan 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 (discussed more 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 (vs. program portability), our intuitive problem solving (vs. rule following), and the sensitivity of our judgments to mood (vs. 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 & 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 does not 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 does not 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, that 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 that 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 that 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, and so forth. 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). An uncontrollable SI would be a very

Related Content