Ethical Guidelines for A Superintelligence

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

October 23, 2014

Abstract

Nick Bostrom, in his new book *SuperIntelligence*, argues that the the creation of an artificial intelligence with human-level intelligence will be followed fairly soon by the existence of an almost omnipotent superintelligence, with consequences that may well be disasterous for humanity. He considers that it is therefore a top priority for mankind to figure out how to imbue such a superintelligence with a sense of morality; however, he considers that this task is very difficult. I discuss a number of flaws in his analysis, particularly the viewpoint that implementing ethical behavior is an especially difficult problem in AI research.

Review of *SuperIntelligence: Paths, Dangers, Strategies* by Nick Bostrom (Oxford U. Press, 2013)

Nick Bostrom, in his new book *SuperIntelligence*, argues that, sooner or later, one way or another, it is very likely that an artificial intelligence (AI) will achieve intelligence comparable to a human. Soon after this has happened — probably within a few years, quite possibly within hours or minutes — the AI will attain a level of intelligence immensely greater than human. There is then a serious danger that the AI will achieve total dominance of earthly society, and bring about nightmarish, apocalyptic changes in human life. Bostrom describes various horrible scenarios and the paths that would lead to them in grisly detail. He expects that the AI might well then turn to large scale interstellar travel and colonize the galaxy and beyond. He argues, therefore, that ensuring that this does not happen must be a top priority for mankind.

The AI need not even have any malicious or megalomaniacal intent. It may just be trying to prove the Riemann hypothesis; but in single-minded pursuit of that goal, it will assemble all the resources, first on earth then in the galaxy, to build additional computational power for that purpose. Or it may have been instructed to make paperclips; in that case, it will turn the whole galaxy into paperclips. Do not think you can escape this doom by instructing it instead to make exactly one million paperclips. If it hears that, it will make the million paperclips, and then exhaust the resources of the universe checking and double checking that it counted correctly.

Bostrom does not at all say *when* he expects this to happen, and, though a committed Bayesian, he does not commit to any probability either. However, the tone of the book suggests that he considers the probability as not less than 1/2.

The first three chapters of Bostrom's book (the "Paths" of his subtitle) are very good. He surveys the different paths that might lead to superhuman intelligence: artificial intelligence, genetic manipulation of humans, brain-computer interfaces, and large networked systems. His discussion of

the state of the art and of the challenges and promise in each direction is well-informed and balanced, though certainly not everyone will agree with all his judgments. Overall this is the best survey of this material that I have seen.

However, it seems to me that there are serious flaws in the discussions of the Dangers and Strategies, which make up the bulk of the book.

The assumption that intelligence is a potentially infinite quantity[1] with a well-defined, one-dimensional value. Bostrom writes differential equations for intelligence, and characterizes their solutions. Certainly, if you asked Bostrom about this, he would say that this is a simplifying assumption made for the sake of making the analysis concrete. The problem is, that if you look at the argument carefully, it depends rather strongly on this idealization, and if you loosen the idealization, important parts of the argument become significantly weaker, such as Bostrom's expectation that the progress from human intelligence to superhuman intelligence will occur quickly.

Of course, there are quantities associated with intelligence that do correspond to this description: The speed of processing, the size of the brain, the size of memory of various kinds. But we do not know the relation of these to intelligence in a qualitative sense. We do not know the relation in brain size to intelligence across animals, because we have no useful measure or even definition of intelligence across animals. And these quantities certainly do not seem to be particularly related to differences in intelligence between people. Bostrom, quoting Eliezer Yudkowsky, points out that the difference between Einstein and the village idiot is tiny as compared to the difference between man and mouse; which is true and important. But that in itself does not justify his conclusion that in the development of AI's it will take much longer to get from mouse to man than from average man to Einstein. For one thing, we know less about those cognitive processes that made Einstein exceptional, than about the cognitive processes that are common to all people, because they are much rarer. Bostrom claims that once you have a machine with the intelligence of a man, you can get a superintelligence just by making the thing faster and bigger. However, all that running faster does is to save you time. If you have two machines A and B and B runs ten times as fast as A, then A can do anything that B can do if you're willing to wait ten times as long.

The assumption that a large gain in intelligence would necessarily entail a correspondingly large increase in power. Bostrom points out that what he calls a comparatively small increase in brain size and complexity resulted in mankind's spectacular gain in physical power. But he ignores the fact that the much larger increase in brain size and complexity that preceded the appearance in man had no such effect. He says that the relation of a supercomputer to man will be like the relation of a man to a mouse, rather than like the relation of Einstein to the rest of us; but what if it is like the relation of an elephant to a mouse?

The assumption that large intelligence entails virtual omnipotence. In Bostrom's scenarios there seems to be essentially no limit to what the superintelligence would be able to do, just by virtue of its superintelligence. It will, in a very short time, develop technological prowess, social abilities, abilities to psychologically manipulate people and so on, incomparably more advanced than what existed before. It can easily resist and outsmart the united efforts of eight billion people who might object to being enslaved or exterminated.

This belief manifests itself most clearly in Bostrom's prophecies of the messianic benefits we will gain if superintelligence works out well. He writes that if a superintelligence were developed, "[r]isks from nature — such as asteroid impacts, supervolcanoes, and natural pandemics — would be virtually eliminated, since super intelligence could deploy countermeasures against most such hazards, or at least demote them to the non-existential category (for instance, via space coloniza-

[1] To be more precise, a quantity potentially bounded only the finite size of the universe and other such cosmological considerations.

tion)". Likewise, the superintelligence, having established an autocracy (a "singleton" in Bostrom's terminology) with itself as boss, would eliminate "risk of wars, technology races, undesirable forms of competition and evolution, and tragedies of the commons."

On a lighter note, Bostrom advocates that philosophers may as well stop thinking about philosophical problems (they should think instead about how to instill ethical principles in AIs) because pretty soon, superintelligent AIs will be able to solve all the problems of philosophy. This prediction seems to me a hair less unlikely than the apocalyptic scenario, but only a hair.

The unwarranted belief that, though achieving intelligence is more or less easy, giving a computer an ethical point of view is really hard.

Bostrom writes about the problem of instilling ethics in computers in a language reminiscent of 1960's era arguments against machine intelligence; how are you going to get something as complicated as intelligence, when all you can do is manipulate registers?

> The definition [of moral terms] must bottom out in the AI's programming language and ultimately in primitives such as machine operators and addresses pointing to the contents of individual memory registers. When one considers the problem from this perspective, one can begin to appreciate the difficulty of the programmer's task.

In the following paragraph he goes on to argue from the complexity of computer vision that instilling ethics is almost hopelessly difficult, without, apparently, noticing that computer vision itself is a central AI problem, which he is assuming is going to be solved. He considers that the problems of instilling ethics into an AI system is "a research challenge worthy of some of the next generation's best mathematical talent".

It seems to me, on the contrary, that developing an understanding of ethics as contemporary humans understand it is actually one of the easier problems facing AI. Moreover, it would be a necessary part, both of aspects of human cognition, such as narrative understanding, and of characteristics that Bostrom attributes to the superintelligent AI. For instance, Bostrom refers to the AI's "social manipulation superpowers". But if an AI is to be a master manipulator, it will need a good understanding of what people consider moral; if it comes across as completely amoral, it will be at a very great disadvantage in manipulating people. There is actually some truth to the idea, central to *The Lord of the Rings* and *Harry Potter*, that in dealing with people, failing to understand their moral standards is a strategic gap. If the AI can understand human morality, it is hard to see what is the technical difficulty in getting it to follow that morality.

Let me suggest the following approach to giving the superintelligent AI an operationally useful definition of minimal standards of ethics that it should follow. You specify a collection of admirable people, now dead. (Dead, because otherwise Bostrom will predict that the AI will manipulate the preferences of the living people.) The AI, of course knows all about them because it has read all their biographies on the web. You then instruct the AI, "Don't do anything that these people would have mostly seriously disapproved of."

This has the following advantages:

- It parallels one of the ways in which people gain a moral sense.
- It is comparatively solidly grounded, and therefore unlikely to have an counterintuitive fixed point.
- It is easily explained to people.

Of course, it is completely impossible until we have an AI with a very powerful understanding; but that is true of all Bostrom's solutions as well. To be clear: I am not proposing that this criterion

should be used as the ethical component of every day decisions; and I am not in the least claiming that this idea is any kind of contribution to the philosophy of ethics. The proposal is that this criterion would work well enough as a *minimal* standard of ethics; if the AI adheres to it, it will not exterminate us, enslave us, etc.

This may not seem adequate to Bostrom, because he is not content with human morality in its current state; he thinks it is important for the AI to use its superintelligence to find a more ultimate morality. That seems to me both unnecessary and very dangerous. It is unnecessary because, as long as the AI follows our morality, it will at least avoid getting horribly out of whack, ethically; it will not exterminate us or enslave us. It is dangerous because it is hard to be sure that it will not lead to consequences that we would reasonably object to. The superintelligence might rationally decide, like the King of Brobdingnag, that we humans are "the most pernicious race of little odious vermin that nature ever suffered to crawl upon the surface of the earth," and that it would do well to exterminate us and replace us with some much more worthy species. However wise this decision, and however strongly dictated by the ultimate true theory of morality, I think we are entitled to object to it, and to do our best to prevent it. I feel safer in the hands of a superintelligence who is guided by 2014 morality, or for that matter by 1700 morality, than in the hands of one that decides to consider the question for itself.

Bostrom considers at length solving the problem of the out-of-control computer by suggesting to the computer that it might actually be living in a simulated universe, and if so, the true powers that be might punish it for making too much mischief. This, of course, is just the belief in a transcendent God who punishes sin, rephrased in language appealing to twenty-first century philosophers. It is open to the traditional objection; namely, even if one grants the existence of God/Simulator, the grounds, either empirical or theoretical, for believing that He punishes sin and rewards virtue are not as strong as one might wish. However, Bostrom considers that the argument might convince the AI, or at least instill enough doubt to stop him in its nefarious plans.

Certainly a general artificial intelligence is potentially dangerous; and once we get anywhere close to it, we should use common sense to make sure that it doesn't get out of hand. The programs that have great physical power, such as those that control the power grid or the nuclear bombs, should be conventional programs whose behavior is very well understood. They should also be protected from sabotage by AI's; but they have to be protected from human sabotage already, and the issues of protection are not very different. One should not write a program that thinks it has a blank check to spend all the resources of the world for any purpose, let alone solving the Riemann hypothesis or making paperclips.

Any machine should have an accessible "off" switch; and in the case of a computer or robot that might have any tendency toward self-preservation, it should have an off switch that it cannot block. However, in the case of computers and robots, this is very easily done, since we are building them. All you need is to place in the internals of the robot, inaccessible to it, a device that, when it receives a specified signal, cuts off the power or, if you want something more dramatic, triggers a small grenade. This can be done in a way that the computer probably cannot find out the details of how the grenade is placed or triggered, and certainly cannot prevent it.

Even so, one might reasonably argue that the dangers involved are so great that we should not risk building a computer with anything close to human intelligence. Something can always go wrong, or some foolish or malicious person might create a superintelligence with no moral sense and with control of its own off switch. I certainly have no objection to imposing restrictions, in the spirit of the Asilomar guidelines for recombinant DNA research, that would halt AI research far short of human intelligence. (Fortunately, it would not be necessary for such restrictions to have any impact on AI research and development any time in the foreseeable future.) It is certainly worth discussing what should be done in that direction. However, Bostrom's claim that we have to accept that quasi-omnipotent superintelligences are part of our future, and that our task is to find a way

to make sure that they guide themselves to moral principles beyond the understanding of our puny intellects, does not seem to me a helpful contribution to that discussion.

Acknowledgements: Many thanks to Andrew Sundstrom for bringing the book to my attention; and to Andrew, Gary Marcus, and Luke Muehlhauser for helpful feedback and enlightening discussions.