Artificial Intelligence 220 (2015) 121–124

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Ethical guidelines for a superintelligence

Ernest Davis

Dept. of Computer Science, New York University, New York, NY 10012, United States

ARTICLE INFO

Article history:
Received 21 October 2014
Accepted 10 December 2014
Available online 30 December 2014

ABSTRACT

Nick Bostrom, in his new book *SuperIntelligence*, argues that 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 disastrous 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.

Nick Bostrom, SuperIntelligence: Paths, Dangers, Strategies, 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.

E-mail address: davise@cs.nyu.edu.

http://dx.doi.org/10.1016/j.artint.2014.12.003
0004-3702/

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 colonization)". 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

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