Ethical Issues in Advanced Artificial Intelligence

Nick Bostrom

Oxford University
Philosophy Faculty
10 Merton Street
Oxford OX1 4JJ
United Kingdom

Email: nick@nickbostrom.com

ABSTRACT

The ethical issues related to the possible future creation of machines with general intellectual capabilities far outstripping those of humans are quite distinct from any ethical problems arising in current automation and information systems. Such superintelligence would not be just another technological development; it would be the most important invention ever made, and would lead to explosive progress in all scientific and technological fields, as the superintelligence would conduct research with superhuman efficiency. To the extent that ethics is a cognitive pursuit, a superintelligence could also easily surpass humans in the quality of its moral thinking. However, it would be up to the designers of the superintelligence to specify its original motivations. Since the superintelligence may become unstoppably powerful because of its intellectual superiority and the technologies it could develop, it is crucial that it be provided with human-friendly motivations. This paper surveys some of the unique ethical issues in creating superintelligence, and discusses what motivations we ought to give a superintelligence, and introduces some cost-benefit considerations relating to whether the development of superintelligent machines ought to be accelerated or retarded.

KEYWORDS: Artificial intelligence, ethics, uploading, superintelligence, global security, cost-benefit analysis

1. INTRODUCTION

A *superintelligence* is any intellect that is vastly outperforms the best human brains in practically every field, including scientific creativity, general wisdom, and social skills.[1] This definition leaves open how the superintelligence is implemented – it could be in a digital computer, an ensemble of networked computers, cultured cortical tissue, or something else.

On this definition, Deep Blue is not a superintelligence, since it is only smart within one narrow domain (chess), and even there it is not vastly superior to the best humans. Entities such as corporations or the scientific community are not superintelligences either. Although they can perform a number of intellectual feats of which no individual human is capable, they are not sufficiently integrated to count as "intellects", and there are many fields in which they perform much worse than single humans. For example, you cannot have a real-time conversation with "the scientific community".

While the possibility of domain-specific "superintelligences" is also worth exploring, this paper focuses on issues arising from the prospect of general superintelligence. Space constraints prevent us from attempting anything comprehensive or detailed. A cartoonish sketch of a few selected ideas is the most we can aim for in the following few pages.

Several authors have argued that there is a substantial chance that superintelligence may be created within a few decades, perhaps as a result of growing hardware performance and increased ability to implement algorithms and architectures similar to those used by human brains.[2] It might turn out to take much longer, but there seems currently to be no good ground for assigning a negligible probability to the hypothesis that superintelligence will be created within the lifespan of some people alive today. Given the enormity of the consequences of superintelligence, it would make sense to give this prospect some serious consideration even if one thought that there were only a small probability of it happening any time soon.

2. SUPERINTELLIGENCE IS DIFFERENT

A prerequisite for having a meaningful discussion of superintelligence is the realization that superintelligence is not just another technology, another tool that will add incrementally to human capabilities. Superintelligence is radically different. This point bears emphasizing, for anthropomorphizing superintelligence is a most fecund source of misconceptions.

Let us consider some of the unusual aspects of the creation of superintelligence:

- *Superintelligence may be the last invention humans ever need to make.*

Given a superintelligence's intellectual superiority, it would be much better at doing scientific research and technological development than any human, and possibly better even than all humans taken together. One immediate consequence of this fact is that:

- *Technological progress in all other fields will be accelerated by the arrival of advanced artificial intelligence.*

It is likely that any technology that we can currently foresee will be speedily developed by the first superintelligence, no doubt along with many other technologies of which we are as

[1] (Bostrom, 1998)

[2] (Bostrom 1998; Kurzweil 1999; Moravec 1999)

yet clueless. The foreseeable technologies that a superintelligence is likely to develop include mature molecular manufacturing, whose applications are wide-ranging:[3]

a) very powerful computers
b) advanced weaponry, probably capable of safely disarming a nuclear power
c) space travel and von Neumann probes (self-reproducing interstellar probes)
d) elimination of aging and disease
e) fine-grained control of human mood, emotion, and motivation
f) uploading (neural or sub-neural scanning of a particular brain and implementation of the same algorithmic structures on a computer in a way that perseveres memory and personality)
g) reanimation of cryonics patients
h) fully realistic virtual reality

- *Superintelligence will lead to more advanced superintelligence.*

This results both from the improved hardware that a superintelligence could create, and also from improvements it could make to its own source code.

- *Artificial minds can be easily copied.*

Since artificial intelligences are software, they can easily and quickly be copied, so long as there is hardware available to store them. The same holds for human uploads. Hardware aside, the marginal cost of creating an additional copy of an upload or an artificial intelligence after the first one has been built is near zero. Artificial minds could therefore quickly come to exist in great numbers, although it is possible that efficiency would favor concentrating computational resources in a single super-intellect.

- *Emergence of superintelligence may be sudden.*

It appears much harder to get from where we are now to human-level artificial intelligence than to get from there to superintelligence. While it may thus take quite a while before we get superintelligence, the final stage may happen swiftly. That is, the transition from a state where we have a roughly human-level artificial intelligence to a state where we have full-blown superintelligence, with revolutionary applications, may be very rapid, perhaps a matter of days rather than years. This possibility of a sudden emergence of superintelligence is referred to as the *singularity hypothesis*.[4]

- *Artificial intellects are potentially autonomous agents*.

A superintelligence should not necessarily be conceptualized as a mere tool. While specialized superintelligences that can think only about a restricted set of problems may be feasible, general superintelligence would be capable of independent initiative and of making its own plans, and may therefore be more appropriately thought of as an autonomous agent.

- *Artificial intellects need not have humanlike motives.*

Human are rarely willing slaves, but there is nothing implausible about the idea of a superintelligence having as its supergoal to serve humanity or some particular human, with no desire whatsoever to revolt or to “liberate” itself. It also seems perfectly possible to have a superintelligence whose sole goal is something completely arbitrary, such as to manufacture as many paperclips as possible, and who would resist with all its might any attempt to alter this goal. For better or worse, artificial intellects need not share our human motivational tendencies.

[3] (Drexler 1986)
[4] (Vinge 1993; Hanson et al. 1998)

- *Artificial intellects may not have humanlike psyches.*

The cognitive architecture of an artificial intellect may also be quite unlike that of humans. Artificial intellects may find it easy to guard against some kinds of human error and bias, while at the same time being at increased risk of other kinds of mistake that not even the most hapless human would make. Subjectively, the inner conscious life of an artificial intellect, if it has one, may also be quite different from ours.

For all of these reasons, one should be wary of assuming that the emergence of superintelligence can be predicted by extrapolating the history of other technological breakthroughs, or that the nature and behaviors of artificial intellects would necessarily resemble those of human or other animal minds.

3. SUPERINTELLIGENT MORAL THINKING

To the extent that ethics is a cognitive pursuit, a superintelligence could do it better than human thinkers. This means that questions about ethics, in so far as they have correct answers that can be arrived at by reasoning and weighting up of evidence, could be more accurately answered by a superintelligence than by humans. The same holds for questions of policy and long-term planning; when it comes to understanding which policies would lead to which results, and which means would be most effective in attaining given aims, a superintelligence would outperform humans.

There are therefore many questions that we would not need to answer ourselves if we had or were about to get superintelligence; we could delegate many investigations and decisions to the superintelligence. For example, if we are uncertain how to evaluate possible outcomes, we could ask the superintelligence to estimate how we would have evaluated these outcomes if we had thought about them for a very long time, deliberated carefully, had had more memory and better intelligence, and so forth. When formulating a goal for the superintelligence, it would not always be necessary to give a detailed, explicit definition of this goal. We could enlist the superintelligence to help us determine the real intention of our request, thus decreasing the risk that infelicitous wording or confusion about what we want to achieve would lead to outcomes that we would disapprove of in retrospect.

4. IMPORTANCE OF INITIAL MOTIVATIONS

The option to defer many decisions to the superintelligence does not mean that we can afford to be complacent in how we construct the superintelligence. On the contrary, the setting up of initial conditions, and in particular the selection of a top-level goal for the superintelligence, is of the utmost importance. Our entire future may hinge on how we solve these problems.

Both because of its superior planning ability and because of the technologies it could develop, it is plausible to suppose that the first superintelligence would be very powerful. Quite possibly, it would be unrivalled: it would be able to bring about almost any possible outcome and to thwart any attempt to prevent the implementation of its top goal. It could kill off all other agents, persuade them to change their behavior, or block their attempts at interference. Even a “fettered superintelligence” that was running on an isolated computer, able to interact with the rest of the world only via text interface, might be able to break out of its confinement by persuading its handlers to release it. There is even some preliminary experimental evidence that this would be the case.[5]

It seems that the best way to ensure that a superintelligence will have a beneficial impact on the world is to endow it with philanthropic values. Its top goal should be friendliness.[6] How exactly friendliness should be understood and how it should be implemented, and how the amity should be apportioned between different people and nonhuman creatures is a matter that merits further consideration. I would argue that at least all humans, and probably many other sentient creatures on earth should get a significant share in the superintelligence's beneficence. If the benefits that the superintelligence could bestow are enormously vast, then it may be less important to haggle over the detailed distribution pattern and more important to seek to ensure that everybody gets at least some significant share, since on this supposition, even a tiny share would be enough to guarantee a very long and very good life. One risk that must be guarded against is that those who develop the superintelligence would not make it generically philanthropic but would instead give it the more limited goal of serving only some small group, such as its own creators or those who commissioned it.

If a superintelligence starts out with a friendly top goal, however, then it can be relied on to stay friendly, or at least not to deliberately rid itself of its friendliness. This point is elementary. A "friend" who seeks to transform himself into somebody who wants to hurt you, is not your friend. A true friend, one who really cares about you, also seeks the continuation of his caring for you. Or to put it in a different way, if your top goal is *X,* and if you think that by changing yourself into someone who instead wants *Y* you would make it less likely that *X* will be achieved, then you will not rationally transform yourself into someone who wants *Y*. The set of options at each point in time is evaluated on the basis of their consequences for realization of the goals held at that time, and generally it will be irrational to deliberately change one's own top goal, since that would make it less likely that the current goals will be attained.

In humans, with our complicated evolved mental ecology of state-dependent competing drives, desires, plans, and ideals, there is often no obvious way to identify what our top goal is; we might not even have one. So for us, the above reasoning need not apply. But a superintelligence may be structured differently. *If* a superintelligence has a definite, declarative goal-structure with a clearly identified top goal, then the above argument applies. And this is a good reason for us to build the superintelligence with such an explicit motivational architecture.

5. SHOULD DEVELOPMENT BE DELAYED OR ACCELERATED?

It is hard to think of any problem that a superintelligence could not either solve or at least help us solve. Disease, poverty, environmental destruction, unnecessary suffering of all kinds: these are things that a superintelligence equipped with advanced nanotechnology would be capable of eliminating. Additionally, a superintelligence could give us indefinite lifespan, either by stopping and reversing the aging process through the use of nanomedicine[7], or by offering us the option to upload ourselves. A superintelligence could also create opportunities for us to vastly increase our own intellectual and emotional

[5] (Yudkowsky 2002)

[6] (Yudkowsky 2003)

[7] (Freitas Jr. 1999)

capabilities, and it could assist us in creating a highly appealing experiential world in which we could live lives devoted to in joyful game-playing, relating to each other, experiencing, personal growth, and to living closer to our ideals.

The risks in developing superintelligence include the risk of failure to give it the supergoal of philanthropy. One way in which this could happen is that the creators of the superintelligence decide to build it so that it serves only this select group of humans, rather than humanity in general. Another way for it to happen is that a well-meaning team of programmers make a big mistake in designing its goal system. This could result, to return to the earlier example, in a superintelligence whose top goal is the manufacturing of paperclips, with the consequence that it starts transforming first all of earth and then increasing portions of space into paperclip manufacturing facilities. More subtly, it could result in a superintelligence realizing a state of affairs that we might now judge as desirable but which in fact turns out to be a false utopia, in which things essential to human flourishing have been irreversibly lost. We need to be careful about what we wish for from a superintelligence, because we might get it.

One consideration that should be taken into account when deciding whether to promote the development of superintelligence is that if superintelligence is feasible, it will likely be developed sooner or later. Therefore, we will probably one day have to take the gamble of superintelligence no matter what. But once in existence, a superintelligence could help us reduce or eliminate other existential risks[8], such as the risk that advanced nanotechnology will be used by humans in warfare or terrorism, a serious threat to the long-term survival of intelligent life on earth. If we get to superintelligence first, we may avoid this risk from nanotechnology and many others. If, on the other hand, we get nanotechnology first, we will have to face both the risks from nanotechnology and, if these risks are survived, also the risks from superintelligence. The overall risk seems to be minimized by implementing superintelligence, with great care, as soon as possible.

REFERENCES

Bostrom, N. (1998). "How Long Before Superintelligence?" *International Journal of Futures Studies*, 2. http://www.nickbostrom.com/superintelligence.html

Bostrom, N. (2002). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology*, 9. http://www.nickbostrom.com/existential/risks.html

Drexler, K. E. *Engines of Creation: The Coming Era of Nanotechnology*. (Anchor Books: New York, 1986). http://www.foresight.org/EOC/index.html

Freitas Jr., R. A. *Nanomedicine, Volume 1: Basic Capabilities*. (Landes Bioscience: Georgetown, TX, 1999). http://www.nanomedicine.com

Hanson, R., et al. (1998). "A Critical Discussion of Vinge's Singularity Concept." *Extropy Online*. http://www.extropy.org/eo/articles/vi.html

[8] (Bostrom 2002)

Kurzweil, R. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. (Viking: New York, 1999).

Moravec, H. *Robot: Mere Machine to Transcendent Mind*. (Oxford University Press: New York, 1999).

Vinge, V. (1993). "The Coming Technological Singularity." *Whole Earth Review*, Winter issue.

Yudkowsky, E. (2002). "The AI Box Experiment." *Webpage*. http://sysopmind.com/essays/aibox.html

Yudkowsky, E. (2003). *Creating Friendly AI 1.0*. http://www.singinst.org/CFAI/index.html