A conversation with Dr. Nick Bostrom, May 4, 2015

Participants

- Dr. Nick Bostrom – Director, Future of Humanity Institute, Oxford Martin School, University of Oxford
- Dr. Nick Beckstead – Research Analyst, the Open Philanthropy Project

Note: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Dr. Bostrom.

Summary

The Open Philanthropy Project spoke with Dr. Nick Bostrom of the Future of Humanity Institute as part of its investigation of global catastrophic risks related to science and technology. Conversation topics included approaches to artificial intelligence safety research, risks and benefits of faster progress in different areas of science and technology, and risks associated with neuroscience and nanotechnology research.

Approaches to artificial intelligence safety

There are different approaches to artificial intelligence (AI) safety, each with its advocates. The field is in its early days, and it is not clear which of these approaches is most promising. The best approach could be one that no one has thought of yet.

At this stage, Dr. Bostrom believes that it makes sense for individuals to pursue the approach they think best. However, the field as a whole should let "many different flowers bloom" by pursuing a variety of approaches.

Research with near-term relevance (e.g. research with implications for near-term advances in AI) will likely be funded in any case. More peculiar problems (e.g. problems that would not have to be confronted until a human-level or superintelligent AI is created) are more likely to be neglected in the short term. Even if these problems were unlikely to exist, it would make sense to focus more philanthropic funding on them, because these are the problems where far-sighted philanthropy is most likely to make a difference.

Promising work that graduate students could undertake

There are a couple of avenues that could contribute to the understanding of how advanced AI systems might behave. Examples include:

- Research outlined in the Machine Intelligence Research Institute's (MIRI's) Technical Research Agenda.
- Paul Christiano's recent work on the structure of approval-directed agents.
- Topics within mainstream computer science, such as:
    - Inverse reinforcement learning
    - Studying how concepts generalize

- Studying how different algorithms would behave if they were run on systems with arbitrary amounts of computing power. What would current algorithms do if there were no constraints on their computing power?

Dr. Bostrom believes that there is currently a lot of value in fluidly exploring the AI safety space, not necessarily by following a rigid research agenda. This exploration could entail engaging in interesting problems and seeing where they lead.

A successful approach to supporting AI safety work might be to focus on picking capable people rather than focus on picking promising projects. Smart people who seem genuinely interested in the problem could drive a lot of progress on AI safety. Drawing talented people to the field, without prescribing what they should do, could be an activity worth supporting (e.g. supporting people to do AI safety work for two years, then evaluating their progress at the end of that period).

Philosophical alignment on AI safety

There is some amount of AI safety awareness that researchers should have in order to contribute to AI safety research, though they may not have to be fully aligned with the underlying philosophy. Ideally, AI researchers would be familiar with the current AI safety literature and discussion, and have an appreciation for the difficulty and importance of the safety problem. AI researchers would also ideally understand the proposed failure modes for superintelligent AI.

It would be useful for new AI researchers to be aware of the body of AI safety literature and to develop new ideas that incorporate an understanding of AI safety. This would be more useful than having current AI researchers rebrand their long-standing work as AI safety research.

Areas where increased research capacity might increase existential risk

Dr. Bostrom is concerned about speeding up technological development in some areas.

Bioengineering

Bioengineering is a broad category. Human cognitive enhancement, cognitive genomics, and gain-of-function research are distinct subfields that might vary a lot in their desirability, from the perspective of reducing existential risk.

Developments in medical technology and human enhancement seem likely to reduce existential risk.

Technologies that enable destructive biological capabilities (i.e. bio-warfare) are very likely to increase existential risk. Technologies that extend our abilities to manipulate the natural environment (e.g. basic bacteria engineering, gene synthesis) are also likely to increase existential risk. It is difficult to tell if there are subareas within these fields might decrease existential risk. Relatedly, attempting to

fund a narrow subfield with where additional work would reduce existential risk might cause other funders to shift their support to other nearby subfields with where additional work would increase existential risk.

AI

Translational AI work

An example of potentially existential risk-reducing work in translational AI is progress in machine translation technology, which might contribute to improving cross-cultural relations.

Systemic improvements

Many grants require that submitted applications be a certain number of pages. Dr. Bostrom is not aware of any studies that suggest that longer grant applications lead to improved funding decisions. It could be that it takes 33% longer to write an application that is twice as long as another, and the shorter application is equally informative, so that 33% of the time spent creating (and evaluating) the application yields no return.

Dr. Bostrom once mentioned this possibility to a representative of a funding agency. The representative replied that their sister agencies have grant application requirements that are twice as long as theirs. The representative explained that agencies use long grant applications to insulate themselves from criticism – they can point to their rigorous application process to defend their decision-making and to show they have done a magnificent job.

Dr. Bostrom is not sure if the field of machine learning is as bogged down with lengthy administrative processes as other fields are. For example, as a field, AI has moved to using technical conferences and conference proceedings as its publishing mechanism. Other fields, like the humanities, continue to use academic journals as their primary publishing mechanism. Academic journal publication is slower than release via conference proceedings or electronic preprint.

Encouraging talented people to do AI research

Dr. Bostrom agrees that adding talented people to the field of AI would be one of the more efficient ways of speeding up progress in AI. However, better than adding talent to the general field of AI would be adding that talent to AI safety research. Also better than adding generally talented people to the field would be recruiting people who care specifically about long-term outcomes, or who are especially conscientious.

If 100 talented people were added to the AI safety field, and 100 added to AI development field, the net effect from an existential risk reduction perspective would be substantially positive, even if acceleration of AI were itself existential risk-negative. This is because the AI safety field is currently much smaller than the AI development field, so there would be an enormous increase in AI safety research capacity, and a moderate increase in AI development research capacity.

Areas where increased research capacity may have a positive impact

Speeding up cognitive genomics might reduce existential risk by contributing to genetic enhancement for intelligence.

It is possible that speeding up the development of surveillance and lie detection technologies would reduce existential risk. Dr. Bostrom is not sure if this is true; it is a question that researchers at FHI currently considering.

A greater legitimization of research related to anti-aging and cryonics fields could encourage people to take long-term outcomes seriously and thereby reduce existential risk. The benefit from legitimization is somewhat separate from the benefit of advances in those fields.

As mentioned above, AI safety work is the subfield within artificial intelligence where it is clearest that faster progress would reduce existential risk.

Areas with positive impacts in the short term

Humane animal agriculture technologies

Humane slaughter, and other technologies that would improve the welfare of animals raised in factory farms, could have large, positive impacts in the short term. For example, genetic engineering that made the animals have better subjective experiences of their condition would be a positive development.

General efforts to ameliorate the conditions of farm animals would also have positive impacts.

Dr. Bostrom hasn't looked closely into the treatment of fish in large fishing operations. However, he wouldn't be surprised if there were new technologies that could fairly inexpensively improve the conditions of fish in these contexts. For example, research into fish stress levels could be used to develop a policy or regulatory proposal aimed at improving the welfare of fish in fishing operations. Another example would be research aimed at improving the slaughtering method – currently fish are usually slaughtered by asphyxiation, which may cause unnecessary suffering. Likewise, there ought to be research into insecticides and technologies for preventing vermin that are less likely to cause suffering than current methods. The number of affected individuals is quite large, so even allowing for some uncertainty as to the mental life of simpler animals, it would still be worth putting some effort into this. This work is unlikely to be undertaken except for ethical reasons, particularly work focused on insects and other very small animals, so there may be low-hanging fruit.

Human enhancement technologies

Mood-enhancing and performance-enhancing drugs for healthy people could have positive short-term impacts. Current regulatory systems appear to use the disease model of medicine, which makes it more difficult to help healthy people. To some extent, drug companies currently produce drugs that can help healthy people. However, this production has to be disguised as treatment for a disease, or disease categories have to be extended. For example, a patient has to be diagnosed with ADHD to be prescribed Ritalin. In this environment, it is hard to make a serious effort towards production and utilization of life-enhancing pharmaceuticals.

Scientifically, it seems fairly simple to produce life-enhancing drugs that are superior to the currently most popular life-enhancing drugs (e.g. alcohol and tobacco). Producing safer, better substitutes for these drugs seems feasible.

Neuroscience

There are multiple paths by which neuroscience research could impact the amount of existential risk:

- Neuroscience research could feed into the development of AI.
- Neuroscience research could enable lie detection, or lead to technologies that can manipulate brains or desires. The effect of this research on existential risk is not clear.

Neuroscience research could have short-term benefits by improving the treatment of mental illness.

Neuroscience is a broad field; specific subfields could have differing impacts on existential risk.

Mapping the connectome

Mapping the connectome of the human brain:

- Might speed up AI development.
- Would be fairly unlikely to lead to whole-brain emulation.
- Might push development in a neuromorphic direction, which may increase existential risk.

Nanotechnology

Existential risks associated with nanotechnology can be divided into three categories

- Nanotechnology research that leads to improved computing hardware, which could contribute to faster development of strong AI.
- The development of nanotech weapons systems, and possible ensuing nanotechnology arms races.
- A "grey goo" scenario, in which someone deliberately creates nanomachines that self-replicate in the natural environment, rather like a

computer virus replicates in a computer network, possibly with catastrophic consequences for the ecosphere.

Dr. Bostrom considers the "grey goo" and arms race scenarios to be the most troubling. In Dr. Bostrom's view, faster progress toward nanotechnology would probably increase existential risk.

All Open Philanthropy Project conversations are available at
http://www.givewell.org/conversations