A conversation with Professor Robin Hanson, April 20, 2016

Participants

- Professor Robin Hanson – Associate Professor of Economics, George Mason University
- Holden Karnofsky – Executive Director, Open Philanthropy Project
- Alexander Berger – Program Officer, U.S. Policy, Open Philanthropy Project

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

Summary

The Open Philanthropy Project spoke with Professor Hanson of George Mason University as part of our investigation into potential risks from advanced artificial intelligence (AI). The conversation focused on reasons why Professor Hanson is skeptical of "hard takeoff" scenarios. Topics included the likely level of complexity of the human brain, typical patterns of improvement in complex systems, and the amount of advantage that new innovations tend to provide in various arenas.

The potential for hard takeoff

Based in part on conversations with AI researchers, Open Philanthropy sees some possibility that a few more key advances in the field of AI, roughly as significant as recent developments in deep learning, may allow researchers to solve most or all remaining problems in developing an AI system that matches or exceeds human intelligence in crucial ways, sufficient to lead to transformative AI.

Professor Hanson summarizes this position as follows:

- Human-level intelligence, as implemented in the human brain, relies on a relatively small number of algorithms and/or large-scale architectural features.
- Therefore, a single new insight into the structure of intelligence could potentially improve AI performance very significantly, and allow subsequent insights to be found more quickly.
- It is feasible that it would require a relatively small number of such insights before human intelligence is achieved or surpassed in many domains.

If the above points are true, some argue that this could lead to a "hard takeoff" scenario, in which a single AI increases its capabilities extremely quickly (e.g. within weeks or days) and is able to obtain a decisive strategic advantage over all other actors.

Key innovations as broadly applicable tools for gaining further abilities

Open Philanthropy finds it more plausible that a small number of insights might in

this way allow an AI to *learn* very quickly how to improve its performance in many areas, rather than that a small number of insights would *directly* result in improved AI performance in many areas. In such a scenario, the eventual extremely capable AI might be quite complicated, but this complexity would have been produced by a relatively small set of learning algorithms or architectures.

Implausibility of hard takeoff due to complexity of intelligence

Professor Hanson believes a hard takeoff scenario is very unlikely, primarily because he disagrees with proponents of hard takeoff about the likely level of complexity of the human brain – i.e., how many different types of tasks it performs, involving how many distinct modules, algorithms, etc.

Professor Hanson believes that the human brain appears to be made up of many different modules with differing structures, and a single cognitive task may involve many modules. The greater the number of different modules or algorithms in a cognitive system, the less of an increase in overall system performance an improvement to any single module is likely to produce.

This is the main reason that Professor Hanson does not share the intuition that a few more innovations would suffice to produce human-level (or greater) intelligence. He also suggests that young AI researchers may tend to be optimistic and predisposed to believe that problems in the field are close to being solved.

Analogy to other complex systems

Complex systems with a large number of parts (e.g. biological, ecological, economic, and software systems) typically improve via many small, accumulated changes to many details, rather than through major, central architectural innovations. Professor Hanson thinks the human brain is most likely similarly complex.

Analogy to innovations by firms in the world economy

New innovations may allow firms to make dramatic gains within a particular industry. However, because innovations tend to have applications only within the relatively narrow scope of a particular industry, an innovation will not allow a firm to outcompete literally every other firm in the world economy. For example, software advances by Google have given it a significant competitive advantage within its own industry, but this advance was still small relative to the whole world economy.

Professor Hanson believes that intelligence is likely fragmented in a way analogous to industries in the world economy, and that a key insight in one area would not immediately give an AI a competitive advantage across many or all arenas.

Ordinary software development as a benchmark

In the absence of specific reasons to expect AI development to be fundamentally different from ordinary software development, the typical characteristics of software advances during the past half-century (e.g. how broad the impacts of a single insight are, how much advantage it gives its developers) should serve as a reference point for how AI development is likely to proceed.

Intuitions on hard takeoff scenarios from economic growth models

Professor Hanson thinks that economic growth theory, while imperfect, is the best conceptual tool currently available for modeling and explaining large-scale growth and change.

Models in economics and the social sciences sometimes include parameters that appear to have the potential, in theory, to influence each other in a way that produces rapid exponential growth. In practice, this rarely happens. Professor Hanson thinks that the intuitions gained from immersion in fields like technology and computer science may make hard takeoff-like scenarios seem more likely, while intuitions developed by immersion in economics and the social sciences make them seem unlikely.

Systemic changes that have produced a dramatic increase in the rate of innovation itself have been, historically, extremely rare. Professor Hanson does not find it implausible in principle that rapid, massive increases in growth rates might occur in the near future, but does think it is very unlikely that this will be caused by the emergence of a single superintelligent AI.

Copying and complementing of new innovations

In the pre-historical era, the first group to acquire a new innovation was sometimes able to gain a decisive advantage, because it was too difficult for other groups to either copy the innovation or complement it through trade (e.g. early humans gaining a decisive advantage over other species). However, for most of human history, innovations have tended to give innovators a significant but not decisive advantage, as other parties have been able to copy or complement innovations relatively quickly (e.g. the initial developments of farming and industry).

The Industrial Revolution

During the Industrial Revolution, Europeans were in general better at gaining new knowledge and producing innovations than other contemporaneous cultures, but because of their need to trade with the rest of the world (since other groups were still superior in other areas), European innovations spread and Europe did not achieve a decisive competitive advantage.

Professor Hanson views new systems of communication between experts (e.g. scientific societies) as the primary cause of the Industrial Revolution, rather than any particular invention (e.g. the steam engine).

Extension to AI development

Professor Hanson thinks the ability of groups to copy or complement the innovations of another group will continue, including between humans and AIs, or between various groups developing AI independently. For this reason, Professor Hanson sees hard takeoff scenarios as requiring the (unlikely) assumption that an AI would not need to cooperate or trade with other actors in any domain in order to accomplish its goals. If the AI did need to trade or cooperate, its ability to gain a unique and decisive advantage would likely be preempted.

Cultural evolution

Joseph Henrich's *The Secret of Our Success* argues that cultural evolution was the primary driver of the emergence of humans as the dominant species (rather than simply the relatively small genetic differences between humans and other primates). Once humans' ability to encode knowledge and behaviors in culture (which allows much greater sharing across generations than genetic inheritance alone) passed a certain threshold, humans were able to advance much more quickly than other species.

Professor Hanson thinks that a community of AIs, on a large enough scale, that are able to diversify and communicate with one another, could potentially advance more quickly due to something akin to cultural evolution, though he points out that computers have already had the ability to easily share knowledge for a long time.