CS546 Adv. ML
Superintelligence Review
Guy Cutting

There is little doubt that artificial intelligence and machine learning are now a permanent feature of the human technology landscape. With the rise of these technologies comes great promise, but also potentially great peril. Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies* admirably contributes to the discussion surrounding the adoption of these technologies. In thoughtful detail, he lays out possible paths to superintelligence, what it might look like and the dynamics of its takeover, and many technical, moral, and political issues that are likely to arise when superintelligence becomes reality. Though the book is (necessarily) largely speculative, it is by no means too soon to begin discussing the potentially revolutionary changes that computational intelligence might bring about in the perhaps very near future. Bostrom's book represents an important effort in the attempt to understand the rapid changes that human society is facing, and to navigate an exciting but very uncertain future.

The book begins with a survey of growth trends across human history, and highlights the recent 'elbow' in that trend. This rapid change in the growth rate of human economy and knowledge suggests, for Bostrom and many other commentators, the possibility of a singularity-like intelligence explosion event. Though the promise of AI has often been overhyped, the possibility of superintelligence is very real. Though programmed AI is often envisioned as the obvious path to this outcome, Bostrom points out that there are other possible paths to this end, including whole-brain emulation and organizations (collective intelligence). In his view, it is not, "a matter of indifference how we get to machine superintelligence. The path taken to get there could make a big difference to the eventual outcome."[1] And there are many "fundamental uncertainties" in how we might get there.

One of them is what superintelligence might look like. Will it be merely a speed improvement in processing, while still qualitatively the same as human intelligence (the speed

[1] Bostrom, Nick. *Superintelligence: Paths, Dangers, and Strategies.* Oxford University Press, 2014. p. 73.

form of superintelligence)? Or will it be a collective form, or a qualitatively different form of intelligence altogether? These questions are difficult to answer with precision, because there is still much that is not understood about intelligence, and technology is advancing so rapidly that the exact contours of future developments are difficult to project. For all the uncertainty, though, Bostrom does a creditable job of contemplating the possible dynamics of the SI explosion - how quickly it will happen, how large a fraction of the world economy will participate[2]. He speculates on the seemingly fantastic idea that the change could happen extremely rapidly (in a matter of hours), a prospect that cannot be ruled out once we consider the possibility of a machine intelligence that can improve itself. Bostrom quantifies the dynamics, providing a valuable framework for discussing the explosion dynamics in more conceptual detail.

There are more questions than answer in the area of future superintelligence. Will there be one, or many?[3] If many, will one have a strategic advantage? Will it be created by a lone hacker, or by a large collaborative research project? Will national authorities see it coming, and will they be able to do anything about it? Unfortunately, world politics seem to be at a precarious juncture - it is hard to imagine a coordinated, international effort to manage the deployment of world-changing AI technology - as the last two decades of rapid advances have shown, the political tends to lag behind the technological. The possibility of a singleton superintelligence looms large, and Bostrom frames the issue in terms of the historical context of the brief American nuclear monopoly. The singleton outcome raises the possibility of an intelligence that could take over the entire ‘cosmic endowment’ of humanity and use it to its ends, whatever those might be.

Bostrom raises a very important distinction in the form of the ‘orthogonality thesis’, the idea that intelligence and motivation are independent variables. He explains that, “Despite the

[2]Bostrom, 88.
[3] Bostrom, 105.

fact that human psychology corresponds to a tiny spot in the space of possible minds, there is a common tendency to project human attributes onto a wide range of alien or artificial cognitive systems“[4]. The discussion of intelligence and values is perhaps the most important part of Bostrom’s work. It is probably possible to build a superintelligence, “that values human welfare, moral goodness, or any other complex purpose its designers might want it to serve” but perhaps equally likely, “and in fact technically a lot easier—to build a superintelligence that places final value on nothing but calculating the decimal expansion of pi.” The challenges of imbuing an artificial intelligence with values are great - not least because questions of ultimate value and goals are perhaps the most contentious that humanity faces, and there is nothing like consensus on these issues. The “value-loading problem”[5] is perhaps the greatest dilemma among the many that are posed by the rise of machine intelligence. A superintelligence could have very non-anthropomorphic goals that are totally at odds with what humanity hopes to achieve, and could pursue them with all of its massive abilities in a malignant way. Bostrom discusses both ‘capability control’ and ‘motivation selection’ methods which might address this issue, and his discussion of these possibilities is valuable. But the fact remains that a machine intelligence could overrun any such attempts to manage its implementation.

Even if an intelligence is not selfish or antagonistic towards humanity, the outcome might be far outside what is intended. A “superintelligence search process might find a solution that is not just unexpected but radically unintended.”[6] Even if such an intelligence was attempting to carry out the wishes of its designers, or of humanity at large, misinterpretation might lead to results which are damaging. There is also a non-trivial possibility that humanity might become enslaved to its own creation - as Bostrom points out, “One area in which superorganisms (or

[4] Bostrom, 147.
[5] Bostrom, 226.
[6] Bostrom, 189.

other digital agents with partially selected motivations) might excel is coercion."[7] The discussion of value-loading is intended to begin to address some of these issues, but it is far from clear that this problem even has a solution. Unfortunately the technical development of AI might outpace any efforts to manage the resulting product by imparting human-like values or goals. Yudkowsky's Coherent Extrapolated Volition (CEV)[8] is a valiant attempt at reckoning with the problem of values and goals in an AI, but it too raises more questions than answers. Though a useful approximation of ultimate value, it might be that humanity is too divided on questions of value for CEV to have any practical meaning. But such ideas are still fruitful areas of discussion. As Bostrom suggests, we will have to harness the tools of decision theory, epistemology, ethics, political organization, and much more in order to manage the enormous challenges presented by machine intelligence.

The promise of artificial intelligence is so great that there is probably no way to slow progress towards its implementation. Even if we had a "macro-structural development accelerator"[9], would we choose to use it to slow progress? Most likely that 'progress' will continue at an accelerating pace. Perhaps humanity is destined to overreach, and perhaps we must learn hard lessons from the consequences of too rapidly developing technology that we do not completely understand. But the growing awareness of these issues is encouraging, and Bostrom's *Superintelligence* makes a valuable contribution to the discussion of where humanity is headed. This discussion is still largely conjectural, but that does not mean we are beginning discussion too early. On the contrary, the issues presented by machine intelligence are so enormous, that any contribution to the discussion surrounding them is of great importance. We

[7] Bostrom, 220.
[8] Bostrom, 256.
[9] Bostrom, 281.

can hope that awareness will grow, and that humanity will navigate the risks posed by such immensely powerful technology.