Let's Read Superintelligence

Josh Snider

April 6, 2019

About this

My blog was originally intended to be a thing where I read programming books and wrote about them. This isn't strictly a programming book, but it's influence warrants a close inspection.

About the author

Nick Bostom is a philosopher at the University of Oxford with a masters in computational neuroscience and a PhD in philosophy. He's most known for his work in existential risks (including superintelligence), the anthropic principle, and human enhancement ethics. That seems like it would give you a decent overview of his background and where he might be coming from for when you read his book.

Fable

I know a lot of people tend to skip prefaces in books, but this is one you should read. You know that owl on the front of the book, it's a metaphor for the titular superintelligence. If you don't mind me spoiling it for you, it basically has a bunch of sparrows deciding that adopting an owl would be a good idea and ignoring the other sparrows who suggest that learning how to raise or domesticate an owl might be a good first step.

Preface

Nick decides that convincing the reader that a superintelligence is imminent is out of scope. I feel that a decent argument for this could have been made and would have given urgency to the rest of the argument, but since Nick spends the rest of the preface talking about how hard the book was to write I'm not going to belabor the point.

Nick, like me and Sandy Mitchell, is also a fan of footnotes which are sometimes funny.

Acknowledgements

This book was part of a public writing process and given that it was published in 2014.

Eliezer Yudkowsky was also listed, which gave me a nice "I-recognize-that-reference" moment.

Contents

The book starts by saying what contemporary AI capabilities are, then talks about how we might get to super-intelligence, spends the bulk of the book talking about how our superintelligence might affect the world, and then wraps up by talking about ways we might affect the superintelligence before it's born. Or at least that's the impression I got from reading the contents before I actually read the book.

Chapter One

This chapter starts by saying it's not about the singularity, then gives a decent overview of the arguments in favor of the singularity, and then states that he only really cares about the part of the singularity that it calls an "intelligence explosion". I, of course, sympathize with wnting to be more focused in how you use language, but that seems a bit odd.

Great expectations

One benefit of not arguing that AGI is imminent is that it doesn't sound weird when he talks about how people keep predicting it to happen in twenty years when they keep being wrong. He also introduces an interesting phrase: "Most technologies that will have a big impact on five or ten years are already in limited use, while technologies that will reshape the world in less than fifteen years probably exist as laboratory prototypes." With that as a framework, I might be willing to put AGI at 15 years from now.

The author also brings up a very good point that historical AGI researches were much less aware of the ethical concerns relating to AI or to the possibility that AI which becomes human-level might become super-human either very quickly or without slowing down at all. One might argue that the increasing interest in these issues is evidence of researches getting closer.

Seasons of hope and despair

AI research really began with the Dartmouth Summer Project. They had some early successes where robots could do a lot of things like solve calculus, see things and move around. Essentially, proving a lot of people who said that a robot could never do X wrong, but these systems weren't particularly extensible and suffered from a combinatorial explosion where things rapidly became impossible as the problem became bigger. As the successes petered out, the field entered an AI winter which was only really ended when computers became powerful enough that evolutionary algorithms and neural networks became workable.

If we take a step back at this moment, we can sense that evolutionary algorithms and neural networks are both approximations of a perfect Bayesian which makes probabilistically optimal use of available data. We know that such an ideal is impossible due to requiring infeasible amounts of computational resources, so we're stuck trying to find something that is good enough. One of the more productive fields for this has been Bayesian networks which are a good way of representing causality and are broadly useful.

State of the Art

Computers at the time of this book's writing were human-level or better at many tasks. Not just a wide variety of games, but also at scheduling logistics, machine translation, and high-frequency trading. Nick brings up a quote from McCarthy about how things that work stop being called AI. This shows a key cultural difference between the present day (2019) and when this book was written (2014). Many companies nowadays are unafraid to call their products "AI" even when they're unlikely to be made into Strong AI.

Opinions about the future of machine intelligence

Current researchers are buoyed in their hopes for the future by the recent commercial and research successes that there work has had. Nonetheless, there's some sentiment that they are less optimistic than previous generations of researchers. There were surveys conducted around this time of experts to get their opinions on when human-like AI would arrive. The consensus seemed to put 10% confidence at 2023, 50% confidence at around 2045, and 90% at around 2080. I'm pretty sure that I've seen these survey results before, so I'd be intererested in seeing the results of another survey now that five years have passed and significant progress has been made in the field.

The survey had two more questions where both Nick and I disagree with the survey results. Given we are both singularatarians it should come as no surprise that we both expect superintelligent AI to come quickly after Human-level AI and that we expect its impact to be more extreme (in either a good or bad way) than the surveyed researches said.

This brings us to the end of the chapter and the conclusion that given that AI research is thought likely to get us a superintelligence this century and that it may have a wide variety of outcomes ranging from human extinction to being very good, we should pay closer attention. One of my own conclusions from reading this book is that it requires a lot of time. The Ciaphas Cain series led me to believe that a 300-page book could be read in an afternoon, but I've taken a comparable amount of time just to read the first chapter of this.

Chapter Two Paths to Superintelligence

TED Talk

https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are