October 11, 2013

Our Final Invention: Is AI the Defining Issue for Humanity?

Humanity today faces incredible threats and opportunities: climate change, nuclear weapons, biotechnology, nanotechnology, and much, much more.

By Seth Baum

Humanity today faces incredible threats and opportunities: climate change, nuclear weapons, biotechnology, nanotechnology, and much, much more. But some people argue that these things are all trumped by one: artificial intelligence (AI). To date, this argument has been confined mainly to science fiction and a small circle of scholars and enthusiasts. Enter documentarian James Barrat, whose new book Our Final Invention states the case for (and against) AI in clear, plain language.

Disclosure: I know Barrat personally. He sent me a free advance copy in hope that I would write a review. The book also cites research of mine. And I am an unpaid Research Advisor to the Machine Intelligence Research Institute, which is discussed heavily in the book. But while I have some incentive to say nice things, I will not be sparing in what (modest) criticism I have.

The central idea is haltingly simple. Intelligence could be the key trait that sets humans apart from other species. We're certainly not the strongest beasts in the jungle, but thanks to our smarts (and our capable hands) we came out on top. Now, our dominance is threatened by creatures of our own creation. Computer scientists may now be in the process of building AI with greater-than-human intelligence ("superintelligence"). Such AI could become so powerful that it would either solve all our problems or kill us all, depending on how it's designed.

Unfortunately, total human extinction or some other evil seems to be the more likely result of superintelligent AI. It's like any great genie-in-a-bottle story: a tale of unintended consequences. Ask a superintelligent AI to make us happy, and it might cram electrodes into the pleasure centers of our brains. Ask it to win at chess, and it converts the galaxy into a supercomputer for calculating moves. This absurd logic holds precisely because the AI lacks our conception of absurdity. Instead, it does exactly what we program it to do. Be careful what you wish for!

It's important to understand the difference between what researchers call narrow and general artificial intelligence (ANI and AGI). ANI is intelligent at one narrow task like playing chess or searching the web, and is increasingly ubiquitous in our world. But ANI can only outsmart humans at that one thing it's good at, so it's not the big transformative concern. That would be AGI, which is intelligent across a broad range of domains – potentially including designing even smarter AGIs. Humans have general intelligence too, but an AGI would probably think much differently than humans, just like a chess computer approaches chess much differently than we do. Right now, no human-level AGI exists, but there is an active AGI research field with its own society, journal, and conference series.

Our Final Invention does an excellent job of explaining these and other technical AI details, all while leading a grand tour of the AI world. This is no dense academic text. Barrat uses clear journalistic prose and a personal touch honed through his years producing documentaries for National Geographic, Discovery, and PBS. The book chronicles his travels interviewing a breadth of leading AI researchers and analysts, interspersed alongside Barrat's own thoughtful commentary. The net result is a rich introduction to AI concepts and characters. Newcomers and experts alike will learn much from it.

The book is especially welcome as a counterpoint to The Singularity Is Near and other works by Ray Kurzweil. Kurzweil is by far the most prominent spokesperson for the potential for AI to transform the world. But while Kurzweil does acknowledge the risks of AI, his overall tone is dangerously optimistic, giving the false impression that all is well and we should proceed apace with AGI and other transformative technologies. Our Final Invention does not make this mistake. Instead, it is unambiguous in its message of concern.

Now, the cautious reader might protest, is AGI really something to be taken seriously? After all, it is essentially never in the news, and most AI researchers don't even worry. (AGI today is a small branch of the broader AI field.) It's easy to imagine this to be a fringe issue only taken seriously by a few gullible eccentrics.

I really wish this was the case. We've got enough other things to worry about. But there is reason to believe otherwise. First, just because something isn't prominent now doesn't mean it never will be. AI today is essentially where climate change was in the 1970's and 1980's. Back then, only a few researchers studied it and expressed concerns. But the trends were discernable then, and today climate change is international headline news.

AI today has its own trends. The clearest is Moore's Law, in which computing power per dollar doubles roughly once every two years. More computing power means AIs can process more information, making them (in some ways) more intelligent. Similar trends exist in everything from software to neuroscience. As with climate change, we can't predict exactly what will happen when, but we do know we're heading towards a world with increasingly sophisticated AI.

Here's where AI can indeed trump issues like climate change. For all its terrors, climate change proceeds slowly. The worst effects will take centuries to kick in. A transformative AI could come within just a few decades, or maybe even ten years. It could render climate change irrelevant.

But AI is not like climate change in one key regard: at least for now, it lacks a scientific consensus. Indeed, most AI researchers dismiss the idea of an AI takeover. Even AGI researchers are divided what will happen and when. This was a core result of a study I conducted of AGI researchers in 2009.

Given the divide, who should we believe? Barrat is convinced that we're headed for trouble. I'm not so sure. AI will inevitably progress, but it might not end up as radically transformative as Barrat and others expect. However, the opposite could also be true too. For all my years thinking about this, I cannot rule out the possibility of some major AI event.

The mere possibility should be enough to give us pause. After all, the stakes couldn't be higher. Even an outside chance of a major AI event is enough to merit serious attention. With AI, the chance is not small. I'd rate this much more probable than, say, a major asteroid impact. If asteroid impact gets some serious attention (by NASA, the B612 Foundation, and others), then AI risk should get a lot more. But it doesn't. I'm hoping Our Final Invention will help change that.

This brings us to the one area where Our Final Invention is unfortunately quite weak: solutions. Most of the book is dedicated to explaining AI concepts and arguing that AI is important. I count only about half a chapter discussing what anyone can actually do about it. This is a regrettable omission. (An Inconvenient Truth suffers the same affliction).

There are two basic types of options available to protect against AI. First, we can design safe AI. This looks to be a massive philosophical and technical challenge, but if it succeeds it could solve many of the world's problems. Unfortunately, as the book points out, dangerous AI is easier and thus likely to come first. Still, AI safety remains an important research area.

Second, we can not design dangerous AI. The book discusses at length the economic and military pressures pushing AI forwards. These pressures would need to be harnessed to avoid dangerous AI. I believe this is possible. After all, it's in no one's interest for humanity to get destroyed. Measures to prevent people from building dangerous AI should be pursued. A ban on high-frequency trading might not be a bad place to start, for a variety of reasons.

What is not an option is to wait until AI gets out of hand and then try mounting a "war of the worlds" campaign against superintelligent AGI. This makes for great cinema, but it's wholly unrealistic. AIs would get too smart and too powerful for us to have any chance against them. (The same holds for alien invasion, though AI is much more likely.) Instead, we need to get it right ahead of time. This is our urgent imperative.

Ultimately, the risk from AI is driven by the humans who design AI, and the humans who sponsor them, and other humans of influence. The best thing about Our Final Invention is that, through its rich interviews, it humanizes the AI sector. Such insight into the people behind the AI issue is nowhere else to be found. The book is meanwhile a clear and compelling introduction to what might (or might not) be the defining issue for humanity. For anyone who cares about pretty much anything, or for those who just like a good science story, the book is well worth reading.