**Our Final Invention: Artificial Intelligence and the End of the Human Era**

 **Book Review**

**Introduction**

In our attempt to create the ultimate intelligence, human beings risk catalyzing their own annihilation. At least according to documentary film maker James Barrat’s 2013 analysis of the dangers of Artificial Intelligence (AI). In his book, “*Our Final Invention*: Artificial Intelligence and the End of the Human Era”, Barrat explores the possible futures associated with the development of artificial superintelligence (ASI); some optimistic and other pessimistic. Barrat falls within the “pessimist” category of AI thinkers but his research leads him to interview and ponder the rosier outlook of some of AI’s most notable researchers. Throughout the book, Barrat finds ample opportunity to debate with optimists such as famed AI researcher Ray Kurzweil and challenges many of their assumptions about humanity’s ability to control an intelligence greater than our own. *Our Final Invention* covers three basic concepts regarding the hazards of AI. First, the inevitability of artificial general intelligence (AGI) and ASI. Second, the rate of change with self-improving intelligence. Third, our inability to understand an intelligence smarter than us and to “hard code” benevolence into it.

Barrat fails to address major challenges underlying the assumptions of most AI researchers. Particularly, those regarding the nature of intelligence and consciousness, AI’s capacity for self-improvement, and the continued exponential growth of computer hardware. These concepts lay at the center of the predictions about our AI future and warrant further investigation.

**Busy Child**

*Our Final Invention* opens with a future scenario where an AGI becomes self-aware and is subsequently unplugged from the internet by its developers as a precaution. It then spends all its resources attempting to manipulate its human keepers into plugging it back into the net where it can replicate itself, interact with the physical world through online orders of goods and services to create whatever it needs to achieve its goals. But what are its goals? Barrat paints a bleak picture of a machine intent on preserving itself and with nanotechnology converts the entire planet as well as you and me into grey goo to be used as computational resources or tools for galactic expansion. Whew. Scary stuff.

**The Inevitability of Artificial Superintelligence**

 In *Our Final Invention*, Barrat spends significant portions of the book (particularly in chapters 3 and 4) exploring the inevitability of AGI and ASI development. Large corporations such as Google, universities, nations, and private stealth companies invest millions of dollars annually into AI research. Whoever cracks the code first on the development of AGI and ASI stands to gain significant advantages across all sectors of the economy, military, and information realms. Barrat highlights the fact that the Defense Advanced Research Projects Agency (DARPA) funds some portion of almost all AI research which gives insight into our own government’s interest in the development of advanced AI. With so many companies and countries interested in achieving AGI and ASI, competition will insight further development even if the risks of AI pose a clear danger to the human species. For example, would the U.S. halt AI research because of public outcry if our government knew that the Chinese were only a couple years away from achieving an artificial general intelligence?

The scope of development needed to produce AGI or ASI is often compared to that of the Manhattan project and the associated implications of the country with the right set of values getting its hands on the technology first. What would the world be like if Nazi Germany had the bomb first? What would the world be like, if China got ASI first? Under no circumstances would the U.S. or any other country for that matter cede this kind of technology to its peers without significant effort. For this reason, many AI researchers believe that the development of AGI and ASI will continue regardless of the risks. Barrat holds that keeping tabs on the rate of AI development becomes a challenge due to the present of “stealth companies” that operate without publicity and may be either small startups with small budgets or backed by large corporations such as Google[[1]](#footnote-1). With so many actors seeking to develop AI, many believe that its development will become inevitable regardless of the public’s understanding of the risks AI may pose to humanity.

**Self-Improving Artificial Intelligence**

 Barrat explores the possibility that future AI will be able change itself to improve its own ability to meet its goals. The fact that AI might have goals outside of those programmed into it or that it may achieve its programmed goals in ways that its inventors never imagined prompted Barrat to interview Steve Omohundro. A PhD physicist from Berkely, Omohundro states that “It’ll be better at making better versions of itself than human programmers could. On the other hand, after a lot of iterations, what does it become”[[2]](#footnote-2). The prospect of self-improving AI both fascinates and frightens many in the AI research field. Genetic programming for example allows bits of code to be evaluated by a computer, mutated (think random changes in the code), and bred with another bit of code to create and entirely new program. Like natural evolution, if you do this process enough times you can get a chihuahua from a dire wolf. This type of iterative improvement results in “code [that] is inscrutable”[[3]](#footnote-3). No longer can a human understand why or even how a piece of code works and it essentially becomes a “black box”. Barrat explains that given the fact that this type of self-improving code development exists today, how much more foreign and inscrutable will and AGI or ASI become after millions of iterations of code improvement.

 Another risk associated with self-improving AI is what experts refer to as the “intelligence explosion”. Mathematician I.J. Good first coined the term and Steve Omohundro asserted that the first AGI will naturally seek self-improvement in order to meet its goals more efficiently. Intelligence explosion generally refers to an AGI’s ability to rapidly grow in capability through recursive self-improvement, possible using something akin to a genetic algorithm or artificial neural network (ANN). ANN’s resemble the way a biological system such as the human brain works to process information through interconnected neurons. The major danger associated with the intelligence explosion lies in the speed at which an AGI might become an ASI before human beings will be able to intervene as referenced above in the busy child scenario.

**Hard Code Benevolence**

 *Our Final Invention* explores the idea of a “friendly” AI to contrast the doom and gloom scenarios that dominate the book. Michael Vassar, president of the Machine Intelligence Research Institute (MIRI), believes that the only responsible way to develop AI is to ensure its “friendliness”. Simply, friendly AI is AI that would do what we would do if we were that smart. An ambitious and worthy goal, if not somewhat ambiguous and filled with assumptions about human nature. Here the challenges refer to our ability to influence an intelligence capable of programming itself. Barrat counters this optimism with questions like "what if the AI finds our restrictions to its behavior too inefficient and simply erases them?" Or perhaps our AI will achieve “friendly” AI goals in ways that we would find abhorrent. More likely, the diversity of human thought, contradictions in human morality, and variety of philosophies which underly our world views would make the notion of friendly AI unachievable without significant alterations in global human societies. My thoughts, not Barrats.

**The Assumptions Under the Hood**

 Barrat rarely mentions the concepts of intelligence and consciousness despite the huge role they play in the development of an intelligence that many believe will surpass that of humanity. An artificial super intelligence may be capable of millions or billions of times more processes than a human per second, but that fact alone may not imply self-awareness or consciousness. The going wisdom presumes that any intelligence as good or greater than our own must, by definition, be all the things that we are such as self-aware and conscious. However, scientific research as well as philosophy has yet to nail down the nature of human consciousness so assumptions about machine consciousness require additional consideration.

 Others question an AI’s ability to improve itself. AI researchers typically hold that since a human can improve its own cognitive capabilities, learn new things and evolve its thinking, then this must be true of advanced machines as well. Not only is the hardware (biological vs mechanical or quantum) vastly different, but the nature of the software that will be used on an AGI or ASI remains unknown (ANN or genetic algorithm for example). The analogy to human improvement may be oversold since human’s engage in simple habit changing or perhaps learning new facts and applying them versus completely rewriting the core program of our consciousness as AI researchers believe an advanced AI would likely do. Other questions remain such as how will an AI retain its original goals as it rewrites itself or will its goals change with each iteration of improved intelligence?

 Historical trends in transistor density have met a fundamental physical barrier and will soon be unable to improve due to phenomenon such as quantum tunneling and uncertainty. Many AI researchers presume advances in computer hardware will continue albeit along perhaps different paths than the traditional transistor size associated with the famous Moore’s Law. Quantum computing continues to improve but requires specialized software built to operate on its specific physical properties. Current AI research may not directly transfer from traditional computers to quantum computers which may present a limitation to AI researchers.

**Conclusion**

James Barrat’s *Our Final Invention* offers a healthy dose of skepticism to those familiar with AI research and the layman alike. He contrasts the world ending busy child scenario with more optimistic views of our future with AI from researchers who envision a eutopia of human and AI symbiosis. He explored the ideas of the intelligence explosion, the inevitability of AI, and the challenges of coding friendliness into a machine smarter than us. I found his observations useful and his scenarios enlightening as we struggle to come to gripes with the realities of machines that can perform and think at levels like that of humans. However, Barrat failed to address the underlying assumptions that drive our perceptions of the future of AI such as the nature of consciousness, presumed exponential growth of computing power and an AI’s ability to improve itself. Barrat’s work will remain a fundamental source of insight into artificial intelligence and those wishing to familiarize themselves with this topic should consider *Our Final Invention* a must read.

1. James Barrat, “*Our Final Invention*: Artificial Intelligence and the End of the Human Era”, Thomas Dunn Books, New York, NY 2013, pg 37-39. [↑](#footnote-ref-1)
2. Ibid. 71. [↑](#footnote-ref-2)
3. Ibid. 75. [↑](#footnote-ref-3)