By Joel Savitz

Originally written for Timothy Rogers' fantastic College Writing II Course at UMass Lowell

Submitted 26 October 2017

Modified for latest publication on 16 March 2019

The Future of Intelligence

Human civilization is at a technological crossroads. Unprecedented advances in industrial and information technology have set in motion a golden age of technological development unconceivable to our human ancestors. The current acceleration of automation has the potential to radically reshape the foundations of the human experience. Researchers across many disciplines are developing systems that can perform tasks previously unique to humans with efficiency far beyond human capacity. This field is commonly known as artificial intelligence. Sophisticated and intelligent automation at a civilizational scale could result in something in between the extremes of a utopian paradigm shift comparable to the industrial revolution in magnitude or a catastrophe resulting in untold horror. This paper will be an exploration into the state of this technology, specifically focusing on potential safety issues. Despite recent advances, artificial intelligence is still in it's infancy and as such, one cannot make substantiated claims about its current or future effects on society with any degree of empirical accuracy or confident integrity. With this in mind, I intend to make a brief survey of safety issues artificial intelligence from several angles, including historical, technical, and speculative.

To begin, I must attempt to define artificial intelligence, which turns out to be a far more difficult undertaking than I had assumed. It is not clear whether such a thing exists or is even possible to create. Even the definition of both “artificial” and “intelligence” alone are not straightforward. The dictionary defines artificial intelligence as “The capability of a machine to imitate intelligent human behavior”(“Artificial Intelligence”). At first glance this may appear to be comprehensive, but problems quickly arise upon closer examination. It is not obvious what “intelligent human behavior” is, if it is even possible to define, and neither is it obvious to detect whether a machine is apparently or truly imitating it. Tegmark notes that artificial intelligence is notorious for having widely varying definitions, and he defines artificial intelligence as being simply “non-biological intelligence,” and intelligence as the “ability to accomplish complex goals” (Tegmark). A machine with this ability, combined with the rapid processing capabilities of computing devices may lead to the creation of a computer that can design computers more sophisticated than itself, and more complex than anything designed by humans. (Bostrom 5). Bostrom calls this as an “intelligence explosion,” and he alludes to an early definition of this concept by mathematician I.J. Good, who worked under Alan Turning to crack the Nazi enigma code:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make. (Bostrom 5)

As the development of artificial intelligence has progressed, Good's views have become more mainstream. Today, it is not outside of the norm to claim that artificial intelligence will greatly surpass human intelligence, but it has not always been this way.

In the early days of the field, the focus was more on disproving skeptics than on preventing an apocalypse. Bostrom notes that early "AI pioneers" hardly even considered safety risks in artificial intelligence (6). They were busy building machines to solve problems that were assumed not to be solvable by computers. Early artificial intelligence research focused on a much more philosophical problem, the issue of constraints. The question has been posed in the realm of human thought as well, and it essentially asks how it is possible to come up with a valid interpretation of the world given that there are nearly infinite interpretations of effectively endless information. The goal was to get a computational machine to construct a functional narrative that can effectively guide it towards a goal. Researchers proposed various frameworks, some suggesting a set "triggering conditions," a programmed response to certain environmental conditions, that would cause the machine to do some processing and improve its response to those same conditions (Holland et al. 9). It took Newell and Simon a decade to put together even a basic theory of human problem solving (Newell and Simon 872). This was in 1972, and both computer science and cognitive psychology were in their relative infancy. At this point, safety considerations were of far lower priority than getting the system to effectively and efficiently solve complex problems.

Artificial intelligence research has come a long way in recent years, and so have the speculations of its most prominent researchers. At a recent conference of these specialists, the median year speculated to be that of the birth of human-level artificial intelligence was 2055

(Tegmark). Of course, the issue of conflicting definitions brings the nature of this speculation. The researchers may have different conceptions of what human-level artificial intelligence would entail. Bostrom, one such specialist, speculates that at some point in the development of superintelligent artificial intelligence, there will be a “crossover” point at which the system’s development becomes driven more by the system itself than by human action (77). This concept is analogous to the above definition of an intelligence explosion. He also speculates that the development of a superintelligent mind would likely be “moderate” or “fast” in pace, as a promising technology such as this will likely gain a lot of support from many significant parties. Tegmark, a founding member of the referenced conference, predicts many possible long-term outcomes for our civilization. Keep in mind that I do not intend to claim that there is any validity to these predictions, nor that they are inevitable. They are simply an interesting thought experiment. Two of his envisioned futures are utopias of either a Libertarian or Egalitarian flavor (Tegmark). In the former, humans live side by side with intelligent machines, both with equal property rights. Tegmark thinks that humans progress will be completely eclipsed by that of the artificially intelligent entities living among them, and that perhaps we would be regarded as second-class citizens in a civilization dominated by hyper-intelligent nonhuman entities. In the latter, all humans live as equals on the backs of intelligent machines that serve them as slaves. This would be a paradise for the humans if the political situation was equally utopian, but it would demand the subordination of non-humans to an underclass position. It is not obvious whether this would be a problem or not, as we do not yet know whether nonhuman subjective experience will emerge from artificially intelligent systems, and whether humans would be able to tell. There is also a possibility of a “zookeeper” outcome, where an artificial intelligence of some kind takes over the planet and confines humans to zoo-like conditions for its own amusement, or perhaps just gets rid of us (Tegmark).

While the long-term possibilities of artificial intelligence strike excitement and fear into the hearts of the average person, the reality of the future is unpredictable. In the present, there are concrete safety hurdles facing the development of artificial intelligence which, if not handled carefully, may result in a dystopian future reality the likes of which science fiction has long speculated. Amodei et al. identified five categories of safety problems currently facing the development of artificial intelligence and machine learning techniques. I will discuss several of the most striking issues and their potential implications.

First, there is the issue of "negative side effects" that may result from the pursuit of a goal (4). Fallible human designers create these systems, and the goal may be either poorly defined or entirely wrong to begin with (5). This could result in an artificially intelligent entity pursuing a goal that has disastrous implications with ruthless efficiency. Next, there is the issue of "reward hacking" (6). This is defined as a robot behaving so as to trigger the function that tells it that it has completed its goal without actually having completed it. An example that they use is with an artificially intelligent agent designed to clean up an office. The agent may either shut off its vision so that it doesn't see any work to be done, thereby completing its goal with impossible speed, or creating more messes on purpose, thereby giving it more work. Both of these scenarios would result in the system activating its reward function more often than intended by the designers. This problem may cause an an AI agent to do more harm than good, negating its value. This is a particularly dangerous issue in the long term, because, "the probability that there is a viable hack affecting the reward function also increases greatly with the complexity of the agent and its available strategies" (8). Like human drug addicts who habitually trigger their natural reward system through the use of psychoactive substances, an artificially intelligent agent with the ability to trigger its reward system by internal fiat "won't be inclined to stop" (9). The nearly automatic anthropomorphization tendency that humans have towards nonhuman entities may be an influence on these safety conclusions. It may be that being human gives us a bias through which we cannot objectively view the potential of nonhuman entities. It is not clear whether any of these problems would even arise in reality. Due to the nature of the field, the most expert opinions are merely educated speculation at best. Despite this, these issues are worth investigating for the potential impact that they may have on the future of humanity.

Another particularly interesting problem is that of "safe exploration," the ability of an agent to search its environment for new inputs and try out new behaviors. The possibility of damage to either the agent of the environment can be mitigated by limiting experimental behavior to certain states that are nearly certain to be safe, but this limits the possibilities for self-improvement (Amodei et al. 14). Humans have their own difficulties in navigating an unfamiliar environment, and the answer to how an artificially intelligent entity could do this efficiently is not obvious. The final problem I will discuss is extremely similar to the last. This is the question of what to do when the observed environment differs sufficiently a familiar environment (16). To eliminate this problem completely requires a generalization of artificial intelligence to all variable environments, a very difficult endeavor. The less "well-specified" the input model is[1] the more difficult it becomes to get an artificially intelligent agent to perform well on all variants of that model. This is why we have been able to teach computers to handle

[1] How easy it is to define the constraints and specifications of possible input in abstract terms that can be represented in a computer model.

abstract concepts and models, like chess, language, and mathematics, while the creation of systems that can successfully navigate through the real world at least as well as humans as been far more difficult.

With an understanding of the current limitations of artificial intelligence in mind, consider this recent breakthrough with a grain of salt. Researchers have produced an agent capable playing and beating any other human or other artificial intelligence agent at the classic chinese board game Go[2] without any prior knowledge. The system learns how to play the game by playing against itself, each time improving its abilities. It went from making completely random moves to a high level understanding of complex human Go concepts (Silver). In the words of Silver, "We've removed the constraints of human knowledge and [the system] is able to create knowledge itself" (Sample). Go is a well-defined model easily abstracted to mathematical form, and this circumvents the last problem raised by Amodei et al. referenced above. This development illustrates the potential for artificial intelligence to rapidly improve itself and surpass human capabilities, demonstrating why safety measures must be of a primary consideration when such a system is applied in a context of potential danger. In another case illustrative of the positive potential of artificial intelligence, Hezaveh et al. managed to create a method to analyze gravitational lenses, which are "complex distortions in spacetime," using neural networks ("Artificial intelligence analyzes"). The researchers were able to take a process that took "up to a few weeks" to complete and required the contributions of specialized experts and created a system that was up to "about ten million times faster" and did not require experts (Hezaveh et al.). This system demonstrates how the generalization of processes that require

[2] A two-player chess-like board game based on simple rules but with highly complex strategy.

highly trained professionals can be democratized and economized, freeing up human capital that can aid the development of society elsewhere. Both of these cases are recent breakthroughs, and it may seem that many of the problems discussed by Amodei et al. are irrelevant to these systems specifically, but they demonstrate that the rate of progress is such that systems with the potential for catastrophic behavior are on the horizon. Researchers are making great strides in the development of self-improving algorithms, and it may not be long before the office cleaning robot envisioned by Amodei et al. is no longer a fantasy. On the other hand, the cleaning robot may remain in in the imaginations of humankind for a long time to come.

Great periods of historical change are best viewed by hindsight. Living in the midst of major turning points in history may seem like regular life for people living at the time, and as humans of the same species living today, we are no different. The rate at which breakthroughs are made in one field after another is absolutely dizzying, but if we are not careful, we may invent the instrument of our demise.

In a worst-case scenario, artificial intelligence may enslave us or exterminate us and the light of human consciousness universe may go dark. In a best-case scenario, we may ascend as a species to unimaginable heights, eliminating the need for self-alienating labor to sustain our standard of living and creating a world of abundance and opportunity. Our current capitalist economy will become an anachronism and resource distribution will have to be reimagined. But, we must exercise caution. In the words of Elon Musk, “If you're not concerned about AI safety, you should be. Vastly more risk than North Korea” (@elonmusk).

But then again, is artificial intelligence independent of any human control even a possibility? Current advances in machine learning allow for large scale analysis of data, but this is a structured input model with finite constraints. An artificially intelligent agent can beat humans at a board game, but this system can be abstractly represented and manipulated as ones and zeros. Artificial intelligence capable of modeling the world inhabited by humans may not even be a possibility. Humans have a hard time agreeing between themselves how to model their interpretations, so why would it be any different for computer systems? Machine learning algorithms have been implicated for creating models that diverge from human models of acceptable behavior. Plomion reports in a forbes article that some algorithms have been found to discriminate on racial or ethnic lines (Plomion). This may be due to the data used to train the algorithms, the programmers of the algorithms, or even unwanted characteristics of the data set that are ignored by human models due to their own biases.

Like all technology, that which currently is being developed under the label of artificial intelligence is only an expansion of human capabilities and not a separate entity in itself. The possibility for the emergence of a truly separate entity is currently unknown. Unfortunately, this limits the definite claims that can be made about the subject, though it remains an area of significance for society at large. The next few decades will be an exciting time for humanity, and there is no doubt that many the technologies covered by the broad label of artificial intelligence will play a major role in the future development of a global civilization. There certainly is a risk to these technologies, but this is an unknown quantity far beyond the scope of my limited research. The only educated speculation I can indulge is that it would be wise to keep a close eye on this field for a long time to come.

Works Cited

Amodei, Dario, et al. “Concrete Problems in AI Safety.” *arXiv:1606.06565v2 [cs.AI]* arxiv.org/abs/1606.06565v2. Accessed 1 October 2017.

“Artificial Intelligence.” Merriam-Webster Dictionary. *Merriam-Webster.* www.merriam-webster.com/dictionary/artificial%20intelligence. Accessed 11 October 2017.

"Artificial intelligence analyzes gravitational lenses 10 million times faster." *Defense & Aerospace Week*, 13 Sept. 2017, p. 66. Global Issues in Context, libraries.state.ma.us/login?gwurl=http://link.galegroup.com/apps/doc/A503953423/GIC?u=mlin_n_umass&xid=b115b409. Accessed 29 Sept. 2017.

Hezaveh, Yashar D, et al. "Fast Automated Analysis Of Strong Gravitational Lenses With Convolutional Neural Networks." *Nature*, vol 548, no. 7669, 2017, pp. 555-557. Springer Nature, doi:10.1038/nature23463.

Holland, John, Keith Holyoak, Richard Nisbett, and Paul Thagard. *Induction*. MIT Press, 1986.

Newell, Allen, and Herbert A. Simon. *Human Problem Solving*. Prentice-Hall, 1972.

Plomion, Ben. "Does Artificial Intelligence Discriminate?." Forbes, 2017, https://www.forbes.com/sites/forbescommunicationscouncil/2017/05/02/does-artificial-intelligence-discriminate/.

Silver, David, et al. “Mastering the Game of Go Without Human Knowledge.” *Nature*. 18 October 2017. doi:10.1038/nature24270. Accessed October 18 2017.

Rikert, Tom. "AI Hype Has Peaked So What'S Next?." TechCrunch. techcrunch.com/2017/09/30/ai-hype-has-peaked-so-whats-next/. Accessed on 2 Oct. 2017.

Sample, Ian. "'It's Able to Create Knowledge Itself': Google Unveils AI That Learns on Its Own." The Guardian, Guardian News and Media, 18 Oct. 2017, www.theguardian.com/science/2017/oct/18/its-able-to-create-knowledge-itself-google-unveils-ai-learns-all-on-its-own.

@elommusk. "If you're not concerned about AI safety, you should be. Vastly more risk than North Korea." *Twitter,* 11 August 2017, 5:29 PM., https://twitter.com/elonmusk/status/896166762361704450