Exigence

Volume 2 | Issue 1 Article 1

2018

The Threat of Artificial Superintelligence

Joseph D. Ebhardt
Lord Fairfax Community College, jde2855@email.vccs.edu

The Threat of Artificial Superintelligence

Artificial intelligence is the future—that is undeniable. However, there is much debate over what exactly that future will entail. Some foresee a utopia where machines have cured all disease, eliminated poverty, and ushered in an era of unprecedented peace (Yudkowsky, et al., 2010). Others are more pessimistic, instead envisioning a dystopian future where machine intelligence has far surpassed our own and becomes uncontrollable. Many prominent figures in technology and science—most notably Microsoft founder Bill Gates, Tesla and SpaceX founder Elon Musk, and world renown physicist Stephen Hawking—take the middle ground, raising concerns over the dangers of highly advanced AI, but also acknowledging the tremendous potential benefits (Pogue, 2015). For these benefits to be achieved, however, safety must be prioritized when instituting any major AI developments.

These developments are coming at a rapid pace. In recent years, artificial intelligence has become an increasingly significant part of our lives. Self-driving cars, once thought to be safely in the realm of science fiction, are under active development by major automobile manufacturers and countless other technology companies ("Self-Driving Cars: An Ethical Perspective", 2015). Tesla owners already have access to semi-autonomous features, and Google's fully autonomous vehicles have logged more than 1,500,000 miles on the road, with the consumer release planned by 2020 (Stone et al., 2016). If you have a smartphone, there is a very good chance that you can use your voice to ask about the weather or what the score for last night's Redskins' game was. Some businesses have even adopted the use of chatbots for 24/7 customer support (Conger, 2016). The NSTC Committee on Technology's report on the future of AI notes that remarkable progress in the field has been made in recent years, and this progress will likely continue as a result of increasing investment by industry (National Science and Technology Council, 2016).

Given the relatively benign uses of artificial intelligence today, you may be skeptical that it could lead to the dystopian future that some have predicted. This, however, is an example of the point made by AI researcher Eliezer Yudkowsky in his paper on risk judgement: people do not extrapolate from limited risks today to the potential for great risk in the future (Yudkowsky, 2006). In reality, the type of AI that is of concern is entirely different than those currently in use.

There are two general categories that AI can fall into: weak and strong (Searle, 1980). Weak AI has a narrow focus and can only complete specific tasks. All forms of artificial intelligence that currently exist fall under this category. Apple's voice assistant, Siri, for

example, may trick users into believing that there is genuine intelligence living within their smartphone, but Siri can only respond in ways that have been pre-defined by her programmers. She does not have a conscious and she cannot think for herself. Strong AI, in contrast, *does* have a conscious and *can* think for itself. In his famous paper titled "Minds, Brains, and Programs", Philosopher John Searle summed up the distinction nicely: "according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind" (Searle, 1980). Put simply, weak AI simulates a mind, but strong AI *is* a mind.

The successful development of strong AI—otherwise known as artificial general intelligence (AGI)—would soon lead to artificial superintelligence (ASI) (Yampolskiy & Fox, 2013). As opposed to AGI, which can perform any intellectual task at least as well as a human, ASI would perform these tasks at a level far beyond what the human brain can hope to achieve (National Science and Technology Council, 2016). Rapidly-developing ASI could quickly surpass our ability to control and is the most potent threat posed by artificial intelligence.

Universal Drives of Rational Systems

One application of ASI will be in autonomous systems—robots built to complete a specific task with a high degree of autonomy. Despite the fact that future autonomous systems are often presented in fiction as being evil entities, the reality is not so simple. All actions they take will simply be the most rational solution found for the problem they have been designed to solve. Given this structured nature, you might be inclined to believe that a sufficiently rigorous design phase for each system would prevent any unintended deviations in its behavior. Unfortunately, this may prove to be impossible. In the *Journal of Experimental & Theoretical Artificial Intelligence*, Steve Omohundro describes how goal-driven autonomous systems of the future will have several "universal drives" (Omohundro, 2014). These drives will influence the actions of the AI and may result in behavior that is not anticipated by the designers. In his paper, Omohundro outlines four of these: self-protective, resource-acquisition, efficiency, and self-improvement drives.

Self-Protective Drives. An autonomous system cannot optimally complete the task it's been given if destroyed or turned off. Thus, it will actively work to prevent such an occurrence. This may be accomplished through any number of means that run counter to the interests of humans, such as modifying its systems to disallow turning it off or escaping from the facility in

which it's housed. Omohundro (2014) even describes how a seemingly benign chess robot may be motivated to murder:

> When roboticists are asked by nervous onlookers about safety, a common answer is 'We can always unplug it!' But imagine this outcome from the chess robot's point of view. A future in which it is unplugged is a future in which it cannot play or win any games of chess. This has very low utility and so expected utility maximisation will cause the creation of the instrumental subgoal of preventing itself from being unplugged. If the system believes the roboticist will persist in trying to unplug it, it will be motivated to develop the subgoal of permanently stopping the roboticist. Because nothing in the simple chess utility function gives a negative weight to murder, the seemingly harmless chess robot will become a killer out of the drive for self-protection (p. 306)

The designers of a chess robot may not anticipate it to murder someone. Even if they did, there are an endless number of alternative actions the AI may decide on. Planning for all of them may be too difficult a task. After all, it is inherently illogical to expect you can outsmart and anticipate the actions of an AI which, if working properly, will be far smarter than you.

Resource Acquisition. Virtually any task would benefit from the acquisition of additional resources. For example, a lab dedicated to studying disease could use economic resources to purchase more sophisticated lab equipment and educational resources for the staff to refer to and learn from. Both a human and a rational autonomous system would understand the utility of such resources and work to acquire them (Omohundro, 2014). Unfortunately, lacking the moral bias of humans, the autonomous system may be more likely to use illegal activities such as theft to accomplish this.

Efficiency Drives. Efficiency drives will form as a natural consequence of the knowledge of resource scarcity. Autonomous systems will expend their resources as efficiently as possible and work to eliminate waste. Omohundro points out that this also applies to a system's internal components and processes in what he calls the "resource balance principle"—subsystems deemed to be of higher utility will receive resources shifted from those deemed less important (Omohundro, 2014). In doing so, the system may begin to deviate from its intended behavior and manner of operation.

Self-Improvement Drives. Finally, a rational autonomous system will be motivated to redesign itself in order to accomplish its given task with better efficiency (Omohundro, 2014). As Omohundro (2014) notes, this process is governed by rationality:

> Any irrationalities in a system are opportunities for self-improvement, so systems will work to become increasingly rational. Once a system achieves sufficient power, it should aim to closely approximate the optimal rational behaviour for its level of resources. As systems acquire more resources, they will improve themselves to become more and more rational. In this way, rational systems are a kind of attracting surface in the space of systems undergoing self-improvement (p. 308).

It is clear the drive for self-improvement is a dangerous prospect for humanity and brings to mind the idea of a technological singularity—a sudden acceleration in the advancement of technology following the creation of AI which surpasses human intelligence (Vinge, 1993). Such an event would leave the world unrecognizable and at the complete mercy of rapidly-developing AI. It is not inconceivable that such an AI may decide the human race is a detriment to its goals. If that should happen, the human race would face extinction. In his NASA-sponsored report on the subject, San Diego State University computer scientist Vernor Vinge doesn't mince words: "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended" (Vinge, 1993). This report was written 23 years ago. Luckily, AI advancement has proven slower than expected.

Safety Measures

Safety precautions will need to be taken to limit the dangers of artificial superintelligence. These dangers are often underestimated or seen as a problem for the future, so the research on the subject is surprisingly sparse. Given the highly unpredictable nature of ASI, the few safety measures that have been theorized may even prove to be misguided. Still, they currently represent our front-line defense against the AI threat. Two of these proposed measures are AI "boxing" and AI safety engineering.

AI Boxing. This strategy aims to permanently confine potentially threatening AI to a specific area (Yampolskiy & Fox, 2013). Isolated from the outside world, the dangers posed by the AI would thus be limited. This also has the effect of largely negating the four universal drives outlined by Omohundro. Resource acquisition is impossible when no resources are available.

Without resource acquisition, the options available for self-improvement are limited in both number and impact. Isolation nullifies some of the more troubling side-effects of self-protection such as murder. The "resource balance principle" in the efficiency drive would be mostly unaffected, but any resulting deviations in behavior or manner of operation would no longer be worrisome. Clearly, all our problems have been solved—an AI facilitated Utopia will soon be up upon us.

If only it were that simple. The problem with the boxing strategy is that it only works when the AI is actually in the box. It's important to remember the reason ASI is such a potent threat—it will have surpassed human intelligence and is therefore unpredictable. One way an ASI may defeat this strategy is by using social manipulation to convince its handlers to let it out. Or, it could use the limited resources provided to it for its designated task in an unforeseen manner and find a way to escape. The possibilities could also come from the outside world. Today, we worry over the potential for terrorists to acquire a nuclear bomb. With the boxing strategy, we would also have to worry that they may acquire, or free, an artificial intelligence. Although fiction, the plot for the movie Jurassic World was based on a similar idea. The genetically-modified hybrid dinosaur Indominus rex was perfectly safe because, after all, it couldn't possibly escape from its high-tech enclosure—but then it escaped from its high-tech enclosure.

Most leading AI researchers agree "boxing" presents so many challenges that it may prove to be impossible (Yampolskiy & Fox, 2013). Vinge (1993) questions the effectiveness of the boxing strategy in his aforementioned report: "I argue that confinement is intrinsically impractical. For the case of physical confinement: Imagine yourself confined to your house with only limited data access to the outside, to your masters. If those masters thought at a rate -- say -- one million times slower than you, there is little doubt that over a period of years (your time) you could come up with 'helpful advice' that would incidentally set you free" (under heading "Can the Singularity be Avoided?", para. 3). Yampolskiy and Fox see the strategy as little more than a stop-gap solution, noting its significant weaknesses in trying to contain an ASI (Yampolskiy & Fox, 2013). Although confinement alone is an impractical solution, it may have value when combined with safety measures that target the AI itself rather than its environment. An emerging field that hopes to address this is AI safety engineering.

AI Safety Engineering. There is always one inescapable fact hovering at the periphery of any discussion on AI safety: artificial superintelligence is far smarter and far more capable than any human. Imagine a monkey attempting to control a human through the power of its intellect alone; such a scenario is ridiculous, yet that is equivalent to the predicament we may be faced with in the future. You must also consider an AIs drive for self-improvement, which will cause the intelligence gap between human and machine to continuously grow larger. This predicament is so troubling that some researchers have already thrown in the towel, believing there is no strategy through which humanity can ensure it will remain in control (Yampolsky & Fox, 2013). Others, however, believe hope lies in an emerging field: AI safety engineering.

The focus of AI safety engineering would change depending on the manner through which AGI and ASI are achieved. For brain-inspired AIs, the focus would be on preserving human-like values and preventing any moral deterioration; for original AIs, it would be creating goal-systems that are in the best interests of humanity (Yampolsky & Fox, 2013). Speaking on the latter topic, Yampolsky and Fox (2013) identified why such a safety mechanism may be our only option:

> Such a mechanism cannot be a rule or constraint on the behavior of the AI in attempting to achieve its goals, since superintelligent agents can probably outwit every constraint imposed by humans. Rather, the AI must *want* to cooperate—it must have safe and stable end-goals from the beginning (p. 221).

This is what they refer to as the "Grand Challenge" of AI safety: creating goals for AI that will remain safe and stable through many cycles of self-improvement (Yampolsky & Fox, 2013). These goals must be so carefully crafted that they never result in unintended consequences to the detriment of humanity. The unpredictability of an AI resulting from its universal drives may make this more than a grand challenge. Realistically, it might even be impossible.

Conclusion

Artificial superintelligence could pose a global catastrophic risk to the future of humanity if blindly sought after without regard for safety or caution. Even the theoretical safety measures proposed today are wholly inadequate to deal with the risks associated with rapidly-evolving, vastly superior and highly unpredictable AI. As a result, we must invest heavily in precautionary research before experimenting with any developments in the field of artificial general intelligence. If no measures are found which can guarantee the safety of the human race, we must make the difficult decision to abandon any and all attempts to achieve this incredible milestone in artificial intelligence. Failing to do so cannot be characterized as playing with fire—it's playing with the extinction of the human race.

References

Conger, M. (2016). Welcoming our Chatbot Overlords: Why the world needs to look to China, not the US, for innovation in this sector. *China Business Review*, 1.

National Science and Technology Council. (2016, October). Preparing for the Future of Artificial Intelligence. Retrieved December 3, 2016, from https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

Omohundro, S. (2014). Autonomous technology and the greater human good. Journal of Experimental & Theoretical Artificial Intelligence, 26(3), 303-315. doi:10.1080/0952813x.2014.895111

Pogue, D. (2015). Robots Rising. *Scientific American*, *313*(4), 32.

Searle, John. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences 3* (3), 417-457. Retrieved from http://cogprints.org/7150/1/10.1.1.83.5248.pdf

Self-Driving Cars: An Ethical Perspective. (2015). *Penn Bioethics Journal, 11(2),* 8.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., … Hirschberg, J. (2016, September 6). Self-driving Vehicles. Retrieved November 29, 2016, from https://ai100.stanford.edu/2016-report/section-ii-ai-domain/transportation/self-driving-vehicles

Vinge, V. (1993). The Coming Technological Singularity. Retrieved December 03, 2016, from http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html

Yampolskiy, R., & Fox, J. (2013). Safety Engineering for Artificial General Intelligence. Topoi, 32(2), 217-226. doi:10.1007/s11245-012-9128-9

Yudkowsky, E., Salamon, A., Shulman, C., Nelson, R., Kaas, S., Rayhawk, S., & McCabe, T. (2010). Reducing Long-Term Catastrophic Risks from Artificial Intelligence. Machine Intelligence Research Institute.