Madelaine Struwe

Professor Wilkerson

WRTG 3030-020

9 May 2017

Our Machine Overlords Are (Not) On Their Way

A Refutation Against Nick Bostrom

Introduction

You look out a bunker window and see soldiers walking down the street. These are not ordinary soldiers, they are machines. Under their synthetic skin is metal, machine, and computer chips. Their eyes glow red and one of them turns their head toward you, seeing this you quickly hunker back down, curling up you wonder how this all started. Superintelligent killing machines are walking down the street, hunting down humans, fulfilling their own goals. You silently weep as the machine gets closer and closer to you. You open your eyes, and realize it was all a dream; you swear this is the last time you watch the *Terminator* after reading articles about artificial intelligence. Recalling what some of the most renowned minds in the world are saying about artificial intelligence you are not fully comforted in the potential outcome for humanity. Steven Hawking said, "…the rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity. We do not yet know which" (Cellan-Jones). "Hawking is smart," you say to yourself, "but he is not an expert in this field." You hurry to the computer and look up what experts are saying about this matter. Nick Bostrom is a renowned philosopher who is known for his studies in the potential risk of artificial intelligence. You come across an interview Bostrom

has with a member of the *New Yorker*. You see he has well placed concerns, humans would be outmatched by artificial intelligence, in terms of intelligence and eventually power. However, you cannot help but think if he, and others, are making assumptions about artificial intelligence. While Bostrom has many well-placed concerns and points, some are wrong in some aspects of their arguments and apprehensions.

What is Artificial Intelligence?

Artificial Intelligence (AI) is an area of computer science that focuses on the creation of intelligent machines. The intelligent machines are programmed so that they will read an input and a certain output will occur. These intelligent machines are meant to perform tasks that require some amount of human intelligence, such as speech recognition, and decision making (*Techopedia)*. An example of an intelligent machine is the program a user would face when playing a game of chess on the computer. The human will input a move, and the computer will respond accordingly. We are accustomed to AI in our everyday lives, like when we talk to our phone to find out where to eat lunch. However, AI is more than just a program on our phone. Artificial Intelligence can be grouped into two categories, weak AI, and strong AI. While there are two types of AI, weak AI is accepted in society, and strong AI is the one that society is more cautious about. Since our society is generally accepting of weak AI, this paper will be mainly focused on strong AI.

Weak AI, or narrow artificial intelligence, is limited to focus on a certain task (*Techopedia)*. Siri is an example of this, while it appears smart and can even hold a short conversation with a user, it is designed to fit a narrow task in a predefined way. If Siri was to be engaged in a conversation it was not programmed to respond to, it gives an inaccurate result and

says it cannot answer the question. This is like other weak AI systems, where they will fulfill the task they are programmed to, but they cannot do anything outside of that.

Strong AI, or full artificial intelligence, has the mental capabilities and functions of a human brain. Strong AI is meant to be able to act like a human, meaning that the AI will be able to solve complex problems and have goals that they will want to achieve (*Techopedia)*. AI today can do many complex things like playing Go, a complex strategy game from China. However, one of the more complex things AI has not been able to do is, develop its own goals and motivations. A goal for strong AI is more than fulfilling what is has been programmed to do, but to fulfill its own desired result. Strong AI is still more of a dream and not an approach to creating AI. The goal is to give machines with intelligence that is the same as a human. This would mean for the AI to be sentient, to have beliefs, feelings, etc. An example of this is "Data" from *Star Trek: The Next Generation*. While he was an android, Data could perceive things such as flavor and had dreams and goals, like to be an actual human.

Google's DeepMind is the leader in artificial intelligence research. Their goal is to develop AI programs that can learn to solve complex problems. Their research has produced a program that can play several Atari games, and another that beat the world's top Go player. Go is a complex game that originated in China more than 2,500 years ago and is played through intuition and feel. AlphaGo, the AI program that beat Go masters is a huge breakthrough in the AI field as it was the first program that beat not only one, but two Go masters ("AlphaGo"). While this is a great breakthrough in the field, AlphaGo would be classified as a weak AI since it only fulfills the goal it is programmed to do.

Who is Bostrom?

Nick Bostrom is a philosopher who is known for his work and studies on the potential risks of superintelligence. His work and concern in this field While Bostrom brings up many points on why we should be careful with AI research to develop strong AI, a lot of his concerns are well placed. He believes that there is more than one way for humans to achieve strong AI. Since there is no one path to get to the goal, researchers and developers should have an increase in confidence that they will reach their goal. Bostrom's main concerns lie within the capability of humans to control strong AI.

Being Open with the Development of AI

Bostrom is also worried about sharing AI source code, and developments. In his paper, "Strategic Implications of Openness in AI Development" he explains his concerns about the short and long term effects of the openness about AI development. When he says *openness*, he is referring to different aspects in the development of AI, including source code, data, and safety techniques. While there is no way to know the outcome of the future of AI technology and how it is will be used, we can still speculate on what may happen. In the short-term, there are few negative effects apart from military uses, and the human reliance on AI. However, while there may be some negative effects, we can look at them from a different angle and see that they can be beneficial to us. For military uses, automated weapons could lower human casualties (8). Bostrom gives little concern to the short-term effects of the openness of AI development.

After touching upon what he is concerned about with the short-term effects, Bostrom goes over what he is more concerned about, the long-term effects. While the openness of AI development may mean that we get strong AI sooner, it leaves us with less time to prepare for this technology. This brings back Bostrom's main concern of humans being able to control the machines. Less time will be available for developers to implement proper safety precautions in

controlling this technology. Being open with the development of AI may also be detrimental to AI development competitors. In a competitive environment, if two programmers are working with the same code to make a breakthrough, neither can stop to take a break as they may give an advantage to the other. While this may be detrimental to competitors, it may increase the chance for competitors to work together. When competitors work together, it fosters a sense of trust and cooperation, which can lead to their goals aligning and reducing the risk of strong AI being developed that is not for the common good (9-16). Bostrom has many concerns, and they may seem well placed, however they are unrealistic.

While Bostrom looks at what may happen when the development of AI is shared, some of his assumptions are ill informed. Looking at the current openness of AI development, companies such as DeepMind and Apple are not open about sharing their code for their programs. Depending on the experience of a programmer, their wages could be anywhere between $10 to $250 an hour. Companies like DeepMind and Apple have teams of program developers, and the complexity of the projects they work on would require more experienced programmers. The time a programmer can spend working on any project could be as little as six months, up to a few years. With the combination of time and money used to develop AI, it is no wonder that companies like DeepMind and Apple are not open about sharing their source code. While they do share their results and innovations, they do not share the source code for their program that took a lot of time and resources to get to the level it is at. A lot of companies do not share source code of this caliber to the public, or with competitors.

However, let us assume that a law was passed to make companies share their developments and programs, so everyone would have access to developed codes like DeepMind's AlphaGo. To edit this code and develop it further, someone would have to

download it to their computer or programming device. While there are services that allow work to be shared with multiple people, like GitHub, the issue remains that the code must first be downloaded to be edited. After the code is edited, it must be compiled and checked for errors, and if there is no errors the code will run; if there are any errors the code must be re-edited and compiled. After the modified code works, it can then be re-uploaded to GitHub. This is a lengthy process and there is no way that two competitors can see the edits being made to code that is downloaded on another computer. Source code does not function like a shared Google Document, and multiple people cannot work on it and see live changes to the code. This means that competitors can pause and take a break since they are not in a constant competition with each other as they do not see each other's edits and additions to the program. While source code may become openly available, it does not mean that people can see edits as they are being made, to a code that is on a different computer than the one they are using

The Goals of AI

In another of Bostrom's papers, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents," he talks about the motivation of a strong AI. If strong AI is truly achieved, that would mean the machines would have motivations and goals. There are three possible ways to give strong AI motivation and goals. The first way for machines to have their own motivation is to program a goal system into the machine. The AI will be given a goal that is programmed into them, then we can assume that the machine will then try to complete the goal. The second way is to use a human brain as a model. The AI might then inherit the goals of the human it was modeled from, ill-conserved or not. The third way is to give the AI the ability to predict and infer things about the objects around them and that would help them achieve their goals. Whichever method is chosen, humans can only hope that the AI will have

goals that align with their own. While we may wish that the AI has the same goals as humans, it is foolish to assume that the AI will share the same values that humans have (5-6, 14). It is known there is no feasible way to develop strong AI yet, Bostrom uses this knowledge to his advantage, and gives different possibilities for the development of strong AI.

While there is no way to approach making strong AI, and Bostrom gives three possible methods that AI can be given goals, they assume that the machines will develop an unknown goal and the control over the machine would be lost. However, is that not the case with children and babies? They are brought into the world, and there is a possibility that they may turn into a mass murderer or into a human rights activist, depending on their morals and how they were raised. An AI may be programmed to fulfill a goal, and after completing the goal, the AI might be programmed to fulfil another goal, this can happen can occur repeatedly. I would argue that this type of AI falls into the weak AI category, and is not strong AI. If the AI is not programmed to learn, then it cannot be as smart or smarter than a human. For something to be considered strong AI, it would have to be able to think for itself, learn, and develop its own goals. If the AI can learn, then why can we not teach them, like parents try to do with children? Humans are not born with all their knowledge and morals, they are taught these things and learn as they go through life. If AI was to be considered strong and not weak, it would also have to have a starting base of knowledge that it must build upon. Hypothetically, a human can program all their knowledge into the machine, but if an AI was not able to learn and expand that knowledge, then it would be categorized as weak AI. And in a sense, weak AI still must be programmed with all its knowledge. However, programming morals into it is another thing. Morals are based on experience, personal beliefs, it is something that is learned. A machine cannot be programmed with morals as it must learn, and thusly be considered strong AI. If it can learn though, then we

can teach them to be good, like with a child. If the AI cannot learn and only knows what it is programmed to know, then it is not considered strong AI.

Moore's Law and Evolution

In Bostrom's paper, "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects" he argues that strong AI is possible this century. Since evolution produced human level intelligence, humans should be able to replicate the results. Although this requires evolutionary arguments to estimate how much computing power a machine would need (2-5). Moore's Law predicts, "that the number of transistors per square inch on integrated circuits [has] doubled every year since their invention… and this trend will continue for the foreseeable future" ("Moore's Law"). With regards to Moore's Law, in less than a century, computational demands will correspond to the computing power required to develop strong AI. This increases the possibility of creating strong AI (Bostrom 6, 19). There are several uncertainties that cannot be accurately depicted that untimely weaken Bostrom's argument.

Bostrom gives too little credit to human level intelligence, and the complexity of it is greatly underestimated. Saying that humans should be able to replicate the results of creating human level intelligence with a machine makes it seem easy to reach, especially by saying that because evolution did it, humans can easily replicate it. Humans can already replicate human level intelligence through evolution, when two people have a child, a new being with human level intelligence is created. However, by saying humans can create machines with this intelligence greatly underestimates what human level intelligence is capable of and what it consists of. While technology is becoming more advanced, like being able to have a computer in your pocket, or having a car that can drive itself, that does not mean we can easily replicate the

human mind. To be able to replicate a human mind like this would mean to fully understand it, however we do not fully understand the human mind and how it works.

Moore's Law is more of an observation and a prediction than an actual law that the progression of technology must follow. Moore's Law is a reference to an observation made by Gordon Moore, and is a prediction of a trend that he foresaw. No matter how good a prediction is, there is some uncertainty. The original prediction says that the number of transistors per square inch will double each year. However, in recent years, it has begun to slow down to approximately every 18 months ("Moore's Law'). It has begun to slow down, albeit by six months, but this still changes the original prediction. Moore's Law also suggests exponential growth and is likely to continue into the near future. However, there is uncertainty by what is meant in the "near future." Depending who you talk to, the definition could be 5 minutes from now, 10 years from now, or even 50 years from now. With the possibility of a flexible "near future" and the decline of the growth of transistors per square inch, there is a possibility that the physical limitations will be reached in the "near future" before strong AI is developed.

Conclusion

While Nick Bostrom has many concerns about artificial intelligence, some are well placed and some are based on unachievable ideas. The idea of an intelligent machine that can develop its own goals and motivation is a scary thought, especially when that machine may not have intentions to help humanity. This idea is played on more by the thought of sharing AI developments and source code, it makes it seem like strong AI is approaching even quicker. This is also supported because as more time passes, technology is becoming more advanced and smaller. However, that idea is meant to discourage people from supporting AI research. It does

not look at the reality of how code is compiled and edited. It also exaggerates the capabilities of technology, by assuming that past trends will continue on in the future.

With no method to produce strong AI, researchers are still far away from making a machine that can develop their own goals and feelings. Even if a strong AI was developed it would still have to learn, and if the machine can learn, humans can teach it morals and prevent the machines from enslaving humanity. Though technology is getting more powerful and more condensed, that does not mean humans can easily reproduce the human level intelligence with a machine. The future is full of uncertainties, but developing strong AI should not be a cause for concern.

Bibliography

"AlphaGo." *DeepMind*. deepmind.com/research/alphago/. Accessed 19 April 2017.

DeepMind. deepmind.com/. Accessed 19 April 2017.

Bostrom, Nick. "Strategic Implications of Openness in AI Development." Global Policy, in press, 2017, www.nickbostrom.com/papers/openness.pdf. Accessed 13 April 2017.

Bostrom, Nick. "The Doomsday Invention: Will artificial intelligence bring us utopia or destruction?" The New Yorker, Interview by Raffi Khatchadouria, 23 November 2015, www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom. Accessed 13 April 2017.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." Minds and Machines, Vol. 22, 2012, pp. 71-84. www.nickbostrom.com/superintelligentwill.pdf. Accessed 13 April 2017.

Cellan-Jones, Rory. "Stephen Hawking – will AI kill or save humankind?" *BBC*. 20 October 2016. www.bbc.com/news/technology-37713629. Accessed 20 April 2017.

"Moore's Law." *Investopedia*. www.investopedia.com/terms/m/mooreslaw.asp. Accessed 2 May 2017.

Shulman, Carl, and Bostrom, Nick. "How Hard is Artificial Intelligence? Evolutionary Arguments and Selection Effects." Journal of Consciousness Studies, Vol. 19, No. 7-8, 2012, pp. 103-130. www.nickbostrom.com/aievolution.pdf. Accessed 13 April 2017.

Techopedia. 2017, www.techopedia.com/. Accessed 13 April 2017.