Rohma Khan

Professor Pete McCabe

English 1012

May 9, 2017

Artificial Intelligence: What Could Possibly Go Wrong

The contents of this paper may read like a science fiction novel. It is not. There are many different types of Artificial Intelligence, or AI. The broadest categories of AI are general AI and narrow AI. Narrow AI, an example of which is Siri, is AI that can do small specific tasks. In Siri's case this is pattern recognition and database searching. General AI is AI that can think. Ray Kurzweil, author of The Singularity Is Near: When Humans Transcend Biology and an expert in the field of AI includes strong AI in general AI. Strong AI, as he describes it, is AI with an aptitude resembling human intelligence, for him particularly pattern recognition and command language. Nick Bostrom, author of Superintelligence: Paths, Dangers, Strategies even further categorizations of superintelligence into 3; speed superintelligence, collective superintelligence and quality superintelligence. Speed superintelligence is AI with the capabilities of human intellect that is just quicker. Collective superintelligence is an AI made of a large number of smaller AI's so that AI's overall performance in many different areas is much better than any current intelligent system. Finally quality superintelligence is a system that is much smarter than the human mind and just as fast. A good example of comparing human intelligence with quality superintelligence is comparing Einstein's intellect to a mouse's. There are two main fields of ethics and morality within AI research.There is machine ethics, which is the field of programming AI with ethics, and robot ethics, which is the duty of developers and rights of

robots. This paper focuses on machine ethics. The future of humanity lies with the actions of AI which depend on the programming of ethics and morals.

In present time there is only narrow Artificial Intelligence. General AI has not yet been created. However, the timeframe made by AI experts for its invention is as early as in 2050 (Armstrong 18). The key idea behind these predictions is this concept of the singularity. The Singularity is described as a point in future period where the change in the rate of technological increase is so rapid that human life is forever transformed. This idea uses the law of accelerating returns, which says the rate of change of our human-created technology is accelerating exponentially and applies it to our future AI capabilities. An example of the law of accelerating returns is Moore's law, which shows that each generation of computer chip doubles its capabilities and speeds every year while becoming smaller. The Singularity also relies on the concept that the better technology gets the more resources will be devoted to its further progression. One method often talked about in the creation of general AI is the emulation of the human brain. However, Paul Allen, cofounder of Microsoft, argues that law of accelerating returns is stopped by something called "the complexity break." This complexity break is that there is not enough understanding of the human brain to emulate it. This argument becomes invalid because there are many routes to create general AI. Some of these routes include nanotubes, molecular computing, self-assembly in nanotube circuits, biological systems emulating circuit assembly, computing with DNA, spintronics (computing with the spin of electrons), computing with light, and quantum computing (Kurzweil 83). While the creation of general AI will eventually happen there are many criticisms of time frame given by experts. In Armstrong's paper "How We're Predicting AI—or Failing To" he shows that the predictions

counteract each other. The time frame of general AI creation is largely irrelevant to this paper, as the focus is on the possible effects of AI on humanity and on the machine ethics. There are two scenarios that general AI will bring. General AI is too intelligent and powerful to have anything else but a massive effect on the world. Either the AI will usher in a new era of humanity or it will destroy us and the choice between the scenarios is based on it's programming.

Strong AI could solve all of the worlds problems. Machine intelligence would constantly be preforming at a higher level of thinking at speeds faster than can be imagined. This thinking would then multiply back. Strong AI would be able to access its own plans and improve them, creating faster and better AI, that would then repeat the cycle but faster. As Kurzweil said "Intelligence, if sufficiently advanced, is, well, smart enough to anticipate and overcome any obstacles that stand in its path" strong AI would have limitless capabilities, things humans would not be able to understand (142). On such capability would be to end world hunger. This could, possibly, be done using cloning technologies and directly cloning animal muscle tissue. By doing this meat could be created at a low cost avoiding hormones and reducing environmental impact with no animals suffering. Another capability of AI would be preserving endangered species and restoring extinct ones. In 2001 scientists synthesized the DNA of the Tasmanian tiger, previously extinct, in attempt to bring it back. Strong AI would be able to. Strong AI possibly would be to send data back in time, as is currently being studied by theoretical physicist Todd Brun of the Institute for Advanced Studies at Princeton (Kurzweil 101). Another capability of AI, this one more likely is the mass production and control of nanotechnology. Nanotechnology is mechanical technology under 100 nanometers. Technology that small is able to rearrange atoms. By doing this is would be possible to create anything from the sum of its parts. The energy cost

of creating materials from one type of material to another would also be eliminated. Nanotechnology could also use superconducting wires to replace aluminum and copper wires to provide greater efficiency. This allows the easier transportation of clean, safe energy as well. There are countless other applications of nanotechnology in the environment that strong AI could come up with. Nanotechnology would also make it possible to augment the human body. Humans would merge with technology. Already humans have the primitive stage of this. Cellphones are a type of augmentation, being connected to the internet through a person's hand, or listening to music through headphones. Nanotechnology allows for much more. Nanotechnology enables humans to redesign their bodies and/or brains. One application of this is a new way of eating. Nanobots in the digestive tract and bloodstream will deliver the nutrients needed. Now, without strong AI, scientist have found the fat insulin receptor gene in mice. They blocked the expression of this gene and found that mice with the blocked gene lived 18% longer, with fewer rates of heart disease and diabetes (Kurzweil 203). Strong AI is capable of working faster and smarter than humans, and look at what has been accomplished without AI. There is so much more possible now. Miguel Nicolelis and Duke Univeristy did research to try to give paralyzed humans a way to control their environment and limbs. They implanted sensors in monkey's brain that allowed them to control a curser on a screen through thoughts alone. The monkey's were able to perfect their control over the robots. (Kurzweil 136) Humans also currently interface cochlear implants, that update on their own. CRISPR is gene editing using nanotechnology that is currently in use now. There are many example's of human's creating technology to do amazing things. Strong Artificial Intelligence will be able to do more. Humans

will be able to enhance the brain using nano-robots, eventually uploading their consciousness, becoming a form of immortal. Strong AI is also what will bring us into space.

All of these technological achievements rely on strong AI being “good”. To ensure this the Foresight Institute has given guidelines to those who create nanotechnology. These defense mechanisms include keeping them in controlled environment, not allowing self-replication, not using materials found in the environment, and more. There are many defense mechanism’s for AI creation as well. One of these defense mechanisms is in the field of robot ethics, and it is the duty of developers to be transparent in any scientific or technological progress. There are ways to get around different defense mechanisms, however.

There are many ways superintelligence could harm as much as it can help humanity. There is the possibility that the first group to create superintelligence will have the frontrunner advantage. The advantage is this, since superintelligence creates new better superintelligence in a faster cycle, the first group to create superintelligence will have all of the power. This will create an AI race that will lead to less defensive strategies being used, and a more dangerous playing ground.

Superintelligence is powerful. As such there must be checks on it in case it decides to go rogue. To assume that superintelligence will share any of the values associated with humans is wrong. Some might say that this is a defense mechanism. That since superintelligence has no inner motivations, it is simply motivated by code, and as such a check is to just program a goal into superintelligence and not to give it unlimited abilities. However, a superintelligence lacking some capabilities could achieve these capabilities, through a variety of loopholes. For example, superintelligence is not programmed to value its own survival. However, superintelligence is

programmed to achieve a goal. Superintelligence will then learn value their own survival because it will help them accomplish their programmed goals. Same is the case with improving their own intelligence or decision making skills. While they may not be programmed to do this, increasing intelligence will help superintelligence to achieve the programmed goal. There are four different types of that improvements that superintelligence will carry out to help improve its chance of achieving its goal; Technological perfection, Resource acquisition, Self-preservation, Goal-content integrity, and, Cognitive enhancement (Bostrom 128). It is therefore prudent that the programming be perfect, as whatever weakness built inside superintelligence it will be able to get rid of through the use of its intelligence. Any superintelligence with any goal becomes an issue because they would have a convergent instrumental reason, an unlimited amount of physical resources and, the ability to eliminate potential threats to itself and its goal system. Human beings might count as potential threats or even physical resources. It then becomes imperative that superintelligence be studied extensively before releasing it to the world. However it cannot be assumed that by simply watching AI while it is in a closed system one can determine if it is a threat. There could have been a treacherous turn or a point it began to trick programmers. Superintelligence does not care what was meant to be programmed in it, only what was. Therefore there could be a perverse instantiation, or a superintelligence discovering some way of fulfilling its final goal that is not the intentions of the programmers who defined the goal.

There are other ways of controlling superintelligence. However the control method must be in place before the program becomes superintelligent. After the program begins o be superintelligence it will have a decisive strategic advantage. It also becomes necessary to implement the solution successfully in the very first system. Capability control methods limit

what the superintelligence can do. There are different methods. The boxing method places the superintelligence in a closed environment where it cannot do harm. One error with boxing method: however, is that places superintelligence in a closed system makes it useless. Another method of control is the incentive method, where they're strong reasons not to engage in harmful behavior, such as rewards built into the superintelligence. However there are problems with this method too. Because it relies on social integration to solve the control problem, the programmer looses control or influence over the superintelligence. It also does not solve the final goal problem. Another control method is stunting or limiting the powers of the superintelligence. However, this is also a problem as this could limit the usefulness of the AI. (Bostrom 146-155)

Instead of controlling AI's capacity there are motivation selection methods. The goal of this is to control AIs through programing. There four types of motivation selection methods. Direct specification is directly creating a set of rules or a goal to be followed. There is the same problem here, though, that is shown earlier. Another selection method is called domesticity which is creating AI that has small, non-ambitious goals. The problem in this is that the language would still be too vague. There is also augmentation, or finding an intelligent being with a moral code in place and making it superintelligent. The problem with this method is that the motivation system of a human being is not very well understood and could get corrupted. The final motivation selection method is indirect normatively setting, or allowing AI to figure out values for itself using some reference given to it.

So far every assumption made in the harm and controlling part of this paper is under the assumption that AI are existing in isolation, but what about AI competing with each other. This would, however, lead to more problems. The original control problem still remains and instead

AI is being thrust into complex social/political/economic movements. In the case of multiple superintelligence interacting there would be massive unemployment. This could be seen as a critique of our current system and not superintelligence, since in most cases robots doing work so humans don't have to is good. However, multiple superintelligent species would result in autonomous weapons and new methods of warfare as different countries fight with each other using AI.

Instead of using defense mechanisms the goal should be to make AI "good". However since AI is intelligent it is not necessarily to program ethics and morality into it. Instead value-loading techniques can be tried. One technique is reinforcement learning. This typically involves creating a reward based system. The problem becomes that as the AI becomes more intelligent it can simply circumvent the values and get the reward. Another technique is value accretion, or experience based values. This is how humans learn morals, however this could be complex and difficult to replicate in a AI. Humans also are not the most moral so basing an all powerful machine on their morals is a bad idea. There is also motivational scaffolding or learning morals one step at a time. The problem with this method is the AI could become too powerful while it still has not fully developed a moral or ethical code. The last method is values learning which simply says that since artificial intelligence is capable of learning, it can learn morals instead of programming them in (Bostrom 202-223). Once the value-loading problem is solved another problem appears. Deciding which values to load becomes the most important question then. Because human morals have changed so often and there is not a consensus on ethics, indirect normatively is used to allow AI to figure out a morally relativist stance. Indirect normatively, as mentioned earlier, is simply to allow superintelligence to decide the values. It is to give the

superintelligence a general outline and allow it to work from there. There are three methods of trying this. The Coherent extrapolated volition method is to tell AI this “Our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted” (Bostrom 228) . This approach is meant to give a source of values and allowing AI to decide what fits, instead of attempting to name every value that is important to humans. Another method is to specifically build AI with the goal of doing what is morally right. These two method are based off superintelligence being smarter than humans and thus being able to better come up with a moral plan. The last method of control in this paper is ratification. That is to make a human ratify every decision and action an AI makes. While these are solutions to individual AI rogue scenarios, In order to ensure that superintelligence has defense mechanisms is to collaborate with people all over the world. This is to avoid the AI race and also to avoid some of the scenarios of government’s taking control of AI for purposes like warfare or conflict.

In “Ethical Guidelines for a Superintelligence” Ernest Davis argues against Bostroms assertions that AI will find loopholes. He says that programmers should not write a program that can spend all the resources of the world for any purpose. However, this argument supports Bostroms claim that an ethical code in the programming is necessary. He also says that any machine should have an off switch that cannot be blocked. This however goes against the premise of artificial intelligence. AI would be able to find the switch and block it.

The precautionary principle states that if there is a small of risk of an action resulting in something terrible then to not to that action. Following this principle general AI development should stop immediately. However, humanity should continue with AI research anyway. Kurzweil argues for "the inevitability of a transformed future," essentially that only under totalitarian relinquishment would advancement in AI be stopped and that even under this totalitarian rule there would black market AI work, which is more dangerous than what is currently happening (274). He also argues that there has always been an "intertwined promise and peril of technological advancement (274)." That is there has always been the promise of peril and it has never stopped technological advancement before. AI could stop world hunger, could cure cancer, could do a lot of good. Is it right to let people suffer of starvation or cancer when AI could solve these problems? Should the current struggle of a person be allowed to continue simply to stop a possible future struggle? Philosophy, morality, ethics and what it means to be human are the core questions surrounding AI, not the questions of if and when it will happen. Currently AI is neither good nor evil. Humanity is, thus, wrestling with the scenarios laid out in this paper, and the concept of the impossible task of codifying our moral and ethical system. In questioning if we should become more than human we must answer what it means to be human and what it means to leave that. We must weigh the risks and possible benefits when deciding how to pursue AI. The only question then left is: Is it worth it?

Works Cited

Allen, Paul G. "Paul Allen: The Singularity Isn't Near." MIT Technology Review. MIT Technology Review, 12 Oct. 2011. Web. 09 Mar. 2017.

Armstrong, Stuart, and Kaj Sotala. 2012. “How We’re Predicting AI—or Failing To.” In Beyond AI: Artificial Dreams, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia.

Bostrom, Nick. Superintelligence: paths, dangers, strategies. New York: Oxford U Press, 2016. Print.

Davis, Ernest. "Ethical guidelines for a superintelligence." Artificial Intelligence 220 (2015): 121-24. Web. 9 Mar. 2017.

Kurzweil, Ray. The singularity is near: when humans transcend biology. New York: Penguin Group, 2005. Print.

"Open Letter on Autonomous Weapons." Future of Life Institute. N.p., n.d. Web. 09 May 2017. <https://futureoflife.org/open-letter-autonomous-weapons/>.