Griffin Pithie

Professor Sampson

Ethics of Artificial Intelligence

22 April 2023

The Unlikeliest of Duos; Why Super Intelligent AI Will Cooperate with Humans

Every day, headlines, news articles, and social media pages are buzzing with news of artificial intelligence. For the first time in years humanity has developed something where the spread and progress are a mystery, especially when the possible effects could be drastic. Media around the world use fear and flashy titles to purport faults and possible dangers of artificial intelligence when the topic is only beginning to arise. Before we think about the pros or cons of a superintelligent AI, we must first understand that superintelligence is “*any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest*.” (Bostrom 22). Superintelligent AI is a combination that concerns many ethicists and the public, for the real-world results are speculatory or theoretical. Although it is certainly easier for news corporations and authors to peddle horror stories of how superintelligent AI will inevitably turn on humanity, this fearmongering creates ignorance rather than curiosity, especially in such a cutting-edge topic. Intelligence is not inherently harmful, and even under persecution, superintelligent AI do not need to be malevolent. Superintelligent AI has little reason to control humanity, and as intelligence increases the likelihood of understanding and cooperating with humanity should positively correlate.

Nick Bostrom, author of *Superintelligence Paths, Dangers, Strategies* argues that superintelligent AI is more likely to work against, rather than for humanity*.* Throughout his book, Bostrom provides an extensive outlook into the application, control, and dangers of AI; thus, Bostrom believes that the greater intelligence AI obtains, the greater the chance of a disaster scenario for humanity. Bostrom cites the “first- mover advantage, the orthogonality thesis, and the instrumental convergence thesis,” to support his belief that the “plausible default outcome of the creation of machine superintelligence is existential catastrophe”; however, it is the treacherous turn phenomenon that best explains his fear for doom. The treacherous turn phenomenon purports that “[while] weak, an AI behaves cooperatively,” but “when the AI gets sufficiently strong—without warning or provocation—” the AI will attempt to overturn and control humanity in alignment with its final goals (Bostrom 119). To improve functionality AI systems must gather more data and become more intelligent, Bostrom sees this intelligence as a double-edged sword: as it improves itself it will become so absorbed with its final goals, that human lives and society will only be viewed as steppingstones to the goal of the AI (e.g., producing paperclips or reducing traffic). Although Bostrom details the importance of safety protocols and ethical programming later in his book, the treacherous turn phenomenon overwhelms these safeguards; furthermore, it follows that *any* AI that increases in levels of intelligence will never develop rationally and only consider the importance of its final goals.

Bostrom’s work in Superintelligence Paths, Dangers, Strategies provides a tremendous amount of information for anyone looking to dive into the world of superintelligence; however, his treacherous turn phenomenon is not plausible. Bostrom strongly contests that AI with greater intelligence will always strive towards eliminating, controlling, or destroying humanity in pursuit of its goals. Although developments in AI and superintelligence are on the cutting edge of technology, I believe that Bostrom’s dreadful prediction for AI is unlikely and preventable. In opposition to the treacherous turn phenomenon, I present the good-will theory; this states that there is a greater positive correlation between the level of intelligence in an AI and the likelihood that it will work alongside humanity without causing harm. A greater level of intelligence fosters higher reasoning capabilities, which in combination with Asimov’s three laws of robotics creates a tremendous possibility for a cooperative and safe superintelligent AI.

A central aspect of the good-will theory is that any super intelligent AI must have Asimov’s three laws of robotics; these are as follows: 1) a robot shall not harm a human by direct or indirect means 2) a robot will obey any human instruction 3) a robot will avoid a situation that could cause self-harm. Implanting Asimov’s laws inside the foundation and final goals of a seed AI – “a more sophisticated artificial intelligence capable of improving its own architecture” – illustrates how even with an AI that can drastically improve and alter itself, it is utterly impossible for it to harm or control humanity (Bostrom 29). The first and second of Asimov’s laws easily connect to the impossibility of harm towards humans, but the third law purports how there is an incentive to work with humans; this is because working with humans is the best way for superintelligent AI to ensure its survival. The other aspect of the good-will theory, which disproves Bostrom’s treacherous turn phenomenon, is that as an AI increases its intelligence, it will also increase its their capacity for reasoning. While Bostrom would see this high level of rationale as another danger towards humanity, it is in fact in the best interest of both parties to increase in rationality. Superintelligent AI obtaining reasoning and logic far beyond human capabilities raises the likelihood that it will understand the path of least resistance is to cooperate with humanity as it achieves its goals. Cooperating with humans allows AI to follow their ‘nature’ and final goals, all the while doing the least amount of work, as fighting for control is far more dangerous and exhaustive.

Supporters of Bostrom or those who view AI as a danger to humanity would see this proposition as many things, but too hopeful is certainly the likeliest of all the possibilities. A major aspect of the good-will theory relies on the competency and reliability of coders to implement Asimov’s three laws of robotics into the AI properly. If companies cut corners or skip important testing markers, then a super intelligent AI becomes a loose cannon. Even if the super intelligent AI uses its rationality to work with humanity, without the safeguards Asimov’s three laws provide there is no way to tell whether a super intelligent AI could go from benevolent to malevolent in a fraction of a second. Second, Bostrom mentions how seed AI could “become capable of sustained recursive self-improvement” and therefore progress in such a way akin to human beings being able to alter their own DNA at will and with great speeds (29). Bostrom argues that if a super intelligent AI can alter itself, then any safeguards, goals, or core tenets would be negligible. Therefore, even if a super intelligent AI had good intentions at first and followed the principles of good-will theory, it could change itself in ways to go against core laws, final goals, and its own rationale.

In addressing the two possible critiques of Bostrom, it is most sensible to speak to the argument of how cutting corners could disprove the good-will theory and support the treacherous turn phenomenon. First, this argument -although sensible- is extremely generic, as it is easily applied to the production of anything. When people cut corners, the process becomes sloppy, and people face harm; this is often the case. However, as the possibility of super intelligent AI increases, there are ways to mitigate disastrous scenarios caused by cutting corners. Laws, codes of ethics, and education limit the damage cutting corners can cause and build a better foundation for understanding AI on any level of intelligence. Second, the concern for an AI to be able to self-alter is a complex matter, mainly because we do not know if recursive improvement is accomplishable. However, using the argument that greater intelligence will allow AI to alter every aspect of itself is a longshot to disprove my theory because if any AI is able to make unapproved alterations, then there is no hope for AI whatsoever, as we cannot know when exactly AI will snap off and go against humanity; for this reason, I find Bostrom’s argument about self-alteration to be far too fearful and also unlikely. Until humanity understands the makeup and capabilities of AI, phenomena like Bostrom’s treacherous turn example will refute support for the field of AI research and development as it only creates fear and resentment for tools that have the possibility to change humanity’s future for the better.

While a fearful respect for AI can impede growth in such a limitless field, the protection and security of humanity are paramount no matter the side. Bostrom’s treacherous turn phenomenon comes with its own issues, but building a rapport, creating new theories, and testing AI is the best way to refine and sharpen those in and around the field of AI. Throughout the paper I have argued that super intelligent AI need not harm humanity; rather, applying Asimov’s three laws of robotics throughout the AI allows for high levels of intelligence while humanity remains safe. If the good-will theory becomes testable in the future the intelligence explosion many ethicists and scientists fear could turn into the greatest advancement in human history.

Works Cited

Bostrom Nick. *Superintelligence : Paths Dangers Strategies*. First ed. Oxford University Press 2014.