Artificial Intelligence: Doing More Harm Than Good?

Shortly before his recent death, renowned and legendary physicist Stephen Hawking delivered an ominous warning about artificial intelligence, claiming it could spell the end of the human race. If you know what is good for you, you would do well to not take such a warning from Stephen Hawking lightly. In fact, as it turns out, his fears are shared by many in the tech community – one of whom is Tesla CEO Elon Musk. Known for his futuristic views, the Tesla CEO is not one to shy away from or fear technology. And yet, Musk has gone to extreme lengths to prevent AI from progressing too far by launching his billion-dollar crusade against it in the form of Neuralink, a company that intends to merge man and machine – or in other words, create cyborgs. With this eccentric venture, Musk seeks to prevent AI from growing smarter (and in turn more dangerous) than humans. What many do not seem to realize, however, is that artificial intelligence has already grown dangerous, has already been to harming people and, consequently, has already begun to pose risks to our society.

It's true - artificial intelligence today is riddled with several flaws that have led to severe consequences. Perhaps one of the most concerning of these flaws is bias – a phenomenon that has only very recently gained media attention in light of the Cambridge Analytica scandal. A data analytics company, Cambridge Analytica unethically harnessed millions of Facebook users' data to (successfully) manipulate the outcome of the last United States election by targeting and playing to individuals' unconscious biases. As it turns out, artificial intelligence can adopt people's biases them when churning out predictions– something it has already been doing.

Since the use of AI algorithms is now so widespread that there remain precious few fields that have yet to be influenced by the technology, this is deeply concerning. One of the benefits of incorporating AI models, be it in healthcare research or approving bank loans, is to increase efficiency and reduce human error. If there are existing biases, however, that error is further proliferated. Perhaps just as concerning is our apparent obliviousness to being manipulated by AI and our blind trust at solutions outputted by these algorithms. As Cathy O'Neil, author of Weapons of Math Destructions, remarks: often people are more willing to trust algorithms and mathematical models because they believe them to be exempt from human bias – which is simply not true.

Take for example the COMPAS system. Created by a for-profit company called Northpointe, COMPAS is able to predict how likely a defendant is to reoffend, and has been used by judges when deciding which inmates to grant parole to. Often, human decision making is inconsistent, faulty and influenced by many seemingly irrelevant factors – and in fact, research shows that judges are more likely to grant parole after lunch break than before. A system like COMPAS, which is not subject to such influencing factors, can therefore be really advantageous when it comes to decision making. Alas, evidence uncovered by Pulitzer Prize winning organization ProPublica has highlighted many dangerous shortcomings with COMPAS and, after uncovering evidence that the algorithm was almost twice as likely to label black inmates as "high risk" (very likely to reoffend) when compared to their white counterparts, argued that the system is riddled with bias. If this is indeed the case, the use of such algorithms does nothing to improve the rampant racial discrimination that is already prevalent in the prison and parole system, and in fact worsens an already bad situation. Instances of biased decision making algorithms are observed in several other scenarios as well. For example, algorithms using online personality test results decide on which individuals are better suited for a particular job, and in other cases they are used to determine who is approved for a loan. These are all very critical decisions that greatly

influence an individual's life, and thus the existence of bias in these decisions is deeply troubling. This begs the question: how does such bias come to be in machine learning algorithms?

Well, algorithms adopt biases as a consequence of the data that they are trained on. By analyzing millions of data points, an algorithm is able to detect patterns and find associations between variables, and if the data is biased so too is the output of the algorithms. For example, if an algorithm is trained on photos of people and the dataset contains many more pictures of Caucasian people than of any other race, the algorithm will have a much harder time recognizing Asian faces. This was certainly true of Nikon's camera algorithms, which misread smiling Asian people for people blinking. Similarly, Hewlett Packard cameras had a harder time recognizing individuals with darker skin tones.

At first glance, this may seem like a simple problem with a simple solution: to use datasets that have been screened and cleaned for bias by individuals before feeding it into algorithms. Unfortunately, it's not nearly that simple. For one, algorithms are reflecting bias that is so deeply ingrained in our data that even top companies like Google are falling prey to it, as we saw when their image recognition algorithm classified black people as gorillas. Further, research by Princeton computer science professor and data privacy expert Arvind Narayanan algorithms linked male names with words like "executive" and female names with words like "marriage". These algorithms, it seems, are susceptible to racial and gender stereotypes which is in fact a reflection of the biases prevalent in our society. Thus, says Narayanan, "it's almost definitional that machine learning is going to pick up and perhaps amplify existing human biases. The issues are inescapable." To complicate things further, bias is often difficult to detect by researchers and data scientists because of our own biases that make it hard to perceive what is biased and what isn't.

As Sandra Wachter - a data ethics researcher at the University of Oxford - puts it: "Algorithms force us to look in a mirror on society as it is". This grows even more necessary and urgent as the "age of algorithm" progresses, as it is unlikely that the implementation of AI and ML algorithms will come to a halt – after all, it is foolish to say that there aren't many advances that ML algorithms have made possible, such as earlier detection of cancerous tumors. Therefore, perhaps what we need to do is stop and take a moment to reevaluate our societies today. Consider this example: an AI created by the company BeautyAI served as a judge for an online beauty pageant after sifting through pictures of all the 60,000 contestants. The results? 43 out of the 44 winning contestants it selected were white. It is not hard to see why this is worrisome – although the algorithms were not designed to take race into account, chances are the training data used reflected the western ideals of beauty. This should force us to reflect and ask ourself introspective questions to reevaluate current standards of beauty that exist today in society. In this case, whilst the results were no doubt disheartening, they were not hugely consequential – but what about when an algorithm is used to select the best candidates for a job, or any other life altering and potentially fatal decisions? For an AI to be fair, it doesn't necessarily need to accurately represent the world as we know it, but rather an alternate world that is not biased – or a utopia of sorts, as Narayanan puts it. But to achieve this, says Narayanan, algorithms need to be able to make judgements of social intelligence such as those that people have debated over for years.

Part of the problem surrounding bias in artificial intelligence algorithms stems from the fact that, in many cases, little is known about how exactly an algorithm works – a phenomenon known as "black box" systems. This was certainly the case with COMPAS. The lack of transparency about a system's inner workings makes it difficult to judge whether or not an

algorithm's results should be trusted. A Wisconsin man sentenced to six years in prison – a decision made by COMPAS – claimed that his rights to due process were violated as little is known about how that decision was reached, but the U.S Supreme Court has declined to review his case. Although we know from ProPublica's assessment of COMPAS that this system is flawed, it seems that did not deter judges from trusting it anyway. Should a system be trusted if information about how a solution was reached is unclear and the data it was trained on is not provided? Experts such as Google's head of AI John Giannandrea caution against this, for we simply do not know enough about the algorithm to assess if its decision making abilities are superior to a human's.

Deep learning is a branch of AI that is gaining popularity amongst practitioners. A method that utilizes layered networks, deep learning algorithms are particularly data hungry and particularly obscure when it comes to how much is known about their inner functioning. In research conducted by Google, engineers ran an image recognition algorithm backwards with hopes of better understanding how it "thinks". What they found was that, like humans, the algorithm narrows in on familiar or more typical features, but the way this functions is significantly different from how human perception operates.

Although explaining AI's inner functioning has proven to be a difficult task thus far, some success has been achieved. A team led by University of Washington professor Carlos Guestrin has been working on a method that enables us to better understand the logic and rationale used by algorithms by having an algorithms return some sort of explanation with the output it generates. The problem with this, however, is that the explanations offered by an algorithm is incomplete. Does this mean that we may never fully understand how AI rationalizes? Perhaps. If this is indeed the case, many argue that ML algorithms should either be employed at a significantly lesser extent, at least until we create AI that we are certain is not making unethical decisions. Perhaps there is a middle ground here somewhere – perhaps, for the time being at least, the best solution would be to exercise caution when it comes to solutions offered by algorithms. If neither an AI or human is able to explain what the AI is thinking, we must err on the side of caution and abstain from employing it.

Lawmakers, for their part, must make greater effort to address these issues of bias and set some standards or guidelines. As well-known philosopher and cognitive scientist Daniel Dennett asks, "What standards do we demand of them [AI], and ourselves?". The Canadian Institute for Advanced Research's pan-Canadian artificial intelligence strategy has been working towards incorporating standards regarding bias in AI, while researchers in the AI space continue to hammer away at different approaches of solving this issue effectively. Further, several organizations - such as the AI Now Institute and Data & Society - have also begun taking action to ensure ethical outcome by urging academics and engineers to work with lawmakers.

Ultimately, we cannot completely stop using AI and, given that the implementation of ML/AI has led to many advancements, nor should we. We do, however, need to come to terms with the fact that we may have to alter how we perceive AI, and perhaps discontinue using it in certain situations. To ensure that AI are not used unethically, it is not unreasonable to suggest that developers and companies using AI should provide governments with documents outlining their goals and what they hope to achieve. Another possibility may be to train teams of third-party individuals to conduct further screening on solutions presented by algorithms to minimize the possibility of causing danger. The bottom line here is that blindly trusting AI could prove more harmful than helpful, and until better solutions are devised, certain precautions must be taken.

References:

Crawford, K. (2016, June 25). Opinion | Artificial Intelligence's White Guy Problem. Retrieved March 20, 2018, from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=1

Hume, K. (2017, December 27). Artificial intelligence is the future-but its not immune to human bias. Retrieved March 20, 2018, from http://www.macleans.ca/opinion/artificial-intelligence-is-the-future-but-its-not-immune-to-human-bias

Knight, W. (2017, May 12). There's a big problem with AI: Even its creators can't explain how it works. Retrieved March 20, 2018, from https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Knight, W. (2017, July 12). Biased algorithms are everywhere, and no one seems to care. Retrieved March 20, 2018, from https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Rutschman, A. S. (2018, March 16). AI gave Stephen Hawking a voice-and he used it to warn us against AI. Retrieved March 20, 2018, from https://qz.com/1231092/ai-gave-stephen-hawking-a-voice-and-he-used-it-to-warn-us-against-ai/

Sharma, K. (2018, February 21). Can We Keep Our Biases from Creeping into AI? Retrieved March 22, 2018, from https://hbr.org/2018/02/can-we-keep-our-biases-from-creeping-into-ai

Simonite, T. (2017, March 06). AI software is better than judges at determining whether criminal defendants are flight risks. Retrieved March 20, 2018, from https://www.technologyreview.com/s/603763/how-to-upgrade-judges-with-machine-learning/

Spielkamp, M. (2017, June 16). We need to shine more light on algorithms so they can help reduce bias, not perpetuate it. Retrieved March 20, 2018, from https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/