Artificial Intelligence: Doing More Harm Than Good?

Not long 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. 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, Musk is not one to shy away from or fear much but artificial intelligence is certainly one of them, with the billionaire going to extreme lengths to prevent it from progressing too far by launching his billion-dollar crusade against it in the form of Neuralink. This is a company that intends to merge man and machine – or in other words, create cyborgs – with the aim of preventing artificial intelligence 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 begun to harm people and has already begun to emerge as a threat 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 and been acknowledged as a relevant factor 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, not only can artificial intelligence play to peoples' biases, but it can also adopt them when churning out predictions– something it has already been doing.

As one may imagine, this is deeply concerning since the use of artificial intelligence algorithms is now so widespread that there remain precious few fields that have yet to be influenced by it, with it affecting decisions ranging from healthcare to bank loans. One of the benefits of implementing such models is to increase efficiency and reduce human error - yet if there are existing biases, that error is further proliferated. Perhaps just as concerning is our apparent obliviousness to being manipulated by artificial intelligence in this way and our blind trust at solutions outputted by algorithms. As Cathy O'Neil, author of Weapons of Math Destructions – a book about bias and AI – 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 system called COMPAS. Created by a company called Northpointe, COMPAS is a system that is able to predict how likely a defendant is to reoffend and is a system that has been used by judges when deciding whom to grant parole. Often, human decision making is inconsistent, faulty and influenced by many seemingly irrelevant factors, as shown by a study which found that judges were more likely to grant parole after just having had a lunch break. Therefore, a system like COMPAS can be really advantageous when it comes to decision making in particular since it is not subject to such influencing factors. Alas, evidence uncovered by Pulitzer Prize winning organization ProPublica has highlighted many faults within COMPAS, a system they argue is riddled with bias after uncovering evidence that the algorithm was almost twice as likely to label black inmates as "high risk", or very likely to reoffend, compared to white inmates. If this is indeed the case, the use of such algorithms does nothing to improve the already rampant racial discrimination that is prevalent in the prison and parole system, but rather 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 algorithms determine who is approved for a loan – all of which are very critical decisions that greatly influence an individual's life, and so 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 as the latter just reflects the bias in the data. For example, if an algorithm is trained on photos of people and the dataset contains many more pictures of Caucasian people than it does 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. However, the problem is 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. In another study, conducted by Princeton computer science professor and data privacy expert Arvind Narayanan, algorithms linked male names with words like "executive" and female names with the word "marriage". Evidently, these algorithms are susceptible to racial and gender stereotypes, and this is so because our society is biased and so naturally our data is as well. 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." Moreover, 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, points out: "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, because it is unlikely that the implementation of artificial intelligence or machine learning algorithms will stop – after all, it is foolish to say that there aren't many, many benefits that these algorithms come with, like detecting cancer for example. Therefore, perhaps what we need to do is stop and take a moment to reevaluate our societies today. Consider this example: an AI was created by the company BeautyAI to judge an online beauty pageant after sifting through pictures of all the 60,000 contestants to choose 44. The results? Most of the chosen winners were white and only one contestant with darker skin was chosen. It is not hard to see why this is worrisome – chances are, training data reflected the western ideals of beauty even though algorithms were not meant to take into account race. This forces us to reflect and ask ourselves introspective questions like "what defines beauty?" or "Is a certain race more beautiful than another?", and reevaluate current standards of beauty that exist today in society. In this case, while the results may have been disheartening, they were not fatal – but what about when an algorithm is used to decide on the best candidates for a job, or any other more potentially life altering decisions? Therefore, in order 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 the COMPAS system. In fact, this lack of knowledge and opaqueness about the systems inner workings makes it difficult to judge whether or not an algorithm's results should be trusted. For example, a Wisconsin man sentenced to six years in prison – a decision made by COMPAS – has claimed that his rights to due process were violated since nothing is known about how that decision was reached, but the U.S Supreme Court declined to review his case. We know after ProPublica's assessment that the output generated by COMPAS is flawed, yet it seems that did not deter judges from trusting them 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 and the reason is simple: we simply do not know enough about the algorithm to assess if it's decision making abilities are superior to a human being's.

One branch of artificial intelligence that we are seeing being used more and more is deep learning, a method utilizing many layered networks that 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 the algorithm narrows in on familiar or more typical features, but the way it functions is vastly different from human perception.

Explaining AI's inner functioning has proven to be a difficult task thus far, but 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. In this research, the algorithm in question returns a few examples of the output with some sort of explanation. However, the problem with this is that the explanations offered by the algorithms are incomplete and fail to explain everything it is doing. Does this mean that we may never fully understand how AI rationalizes? Perhaps. If that is indeed the case, then many argue that we may either have to stop using algorithms as much or come to fully trust the algorithm's judgement that needs to take into account our society and social norms to ensure that AI is not making unethical decisions. However, maybe 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, and if neither an AI or human is able to explain what the AI is thinking, then we mustn't use it for fear that it does more harm than good.

Moreover, lawmakers 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, whilst researchers continue to work on different approached to solve the issue in a more effective and complete way. Further, several organizations, such as the AI Now Institute and Data & Society, have now begun taking action as well by urging academics and engineers to work in correspondence with lawmakers to discuss how best to tackle this situation and ensure ethical outcomes.

Ultimately, we cannot completely stop using artificial intelligence – and nor should we, because there is no doubt that it has led to and will continue to lead to many improvements and advancements. However, we need to come to terms with the fact that we may have to change how we perceive AI, and prevent it from being used in some situations. It is not unreasonable to

suggest that developers and companies using AI should first outline and explain their goals and what they aim to achieve by implementing artificial intelligence algorithms and be ready to provide these documents to governments, just to ensure that they are not intended to be used to cause harm. Perhaps another possibility would be to train teams of individuals to conduct further screening on the solutions presented by an algorithm to minimize the possibility of causing danger. The bottom line is that blindly trusting AI could prove more harmful than helpful, and until better solutions are devised (if at all), certain precautions must be implemented.

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/