IDC4UM: Philosophy paper

# Bias in Bits: Unravelling the Ethical Dilemma of Artificial Intelligence's Advancement and its repercussions on marginalized communities

Guiding Question: Should the access of AI models be controlled and regulated for the benefit of humanity?

IDC4UM-01

April 12th, 2024

Carrie Chen

Mr. Lewis

Word count: 1117

Artificial Intelligence (AI) in the 21st century has rapidly developed into a second wave of automation with the potential to wipe out entire industries at an international level. Models like Cognition's Devin have over 4.3 billion neurons and 13 million parameters to achieve a 13.86% resolve rate on GitHub's issues end-to-end board, almost 7x the previous state of the art 1.96% ("Introducing Devin, the first AI software engineer"). Many researchers argue that the AI must be completely unregulated to ensure its advancement is not tainted by political intervention (Christian #43). However, as the sense of individuality of these model's seem to grow, it is paramount to note that from an instrumental standpoint, they're simply tools employed by an individual; therefore the ethics towards their use rest in those employing the machines. By examining cases of unrestrained AI models and their repercussions on already marginalized communities, it is imperative that, just like any other resource, governments must actively surveil and regulate them to prevent misuse.

For decades, AI has been the momentum for high-level STEM engineers across the world in hopes of bringing accessible, open-sourced solutions to computationally intensive problems, fostering innovation and democratizing technological advancement. While the open-sourced nature of AI has undoubtedly facilitated collaboration and innovation, it has also raised concerns about the potential for harm, particularly when AI models perpetuate and exacerbate existing biases. In Albaroudi et al.'s paper, "A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring," more than 200,000 job applications were revised using a formal meta-analysis of 97 field experiments. The study's findings revealed the existence of pervasive discrimination against BIPOC communities in popular AI hiring algorithms due to a skewness in their data-sets. Without proper data collection guidelines, AI models will unintentionally favour applicants who align with the demographic trends observed in their training set. Furthermore, as institutions begin incorporating algorithms to automate high risk decisions such as theft identification or whether a defendant should be granted bail, the potential for life changing mistakes to occur increases. In 2016, a team of data journalists decided to take a closer look at a model used by parole boards across the United States to grant or deny parole called COMPAS (Christian #4). It's purpose was to assign algorithmic risk scores to prisoners on a scale from 1 to 10 to determine their eligibility for parole. Despite controlling for various factors such as age, gender, and criminal history, the analysis found that African American defendants were more likely to be assigned higher risk scores than white defendants for similar offences (Christian #4). COMPAS was not the exception, rather, gender, sexuality, and race based biases was the norm in algorithms trained on human datasets ("Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies" #3). Although the tech is alluring, it is clear that it's efficiency was made at the expense of the human condition, thus strict guidelines for the use of AI in human-focused interactions must be set.

Jean-Jaques Rousseau noted in 1762 that the relationship between a nation and its citizen is social contract of mutual obligation, wherein individuals relinquish certain freedoms in exchange for protection and order provided by the state (Rousseau #27). In the context of AI, the government has an obligation to ensure that AI technologies are developed in a manner that upholds existing societal values, protects individual rights and promotes the common good because their citizens trust them to do so. The current absence of enforceable regulatory framework surrounding the applications of AI pose a threat to the Social Contract as there's no assignment of liability in cases of harm. In Song et. al's paper, "Automated Vehicle Crash Sequences: Patterns and Potential," it was reported that self-driving vehicles were more than twice as likely as traditional vehicles to become involved in auto accidents ("Automated Vehicle Crash Sequences: Patterns and Potential" #2). More precisely,  for every crash in conventional vehicles over a million vehicle miles, there were 2.187 crashes in Autonomous Vehicles (AV) ("Automated Vehicle Crash Sequences: Patterns and Potential" #6). Despite the statistics, there are no federal guidelines for who is to blame in case of an AV accident, nor are there any for gender bias in hiring algorithms, of race based discrimination in parole selection etc… These uncertainties not only hinder innovation and progress but also threaten fundamental principles of fairness, accountability, and social cohesion. There should be ethical standards put in place on the performance of these models before they're used on real people to ensure situations like COMPAS do not repeat in the future. By enacting regulatory frameworks, the government can fulfill its role as the guardian of the social contract, ensuring that AI serves the interests of society as a whole rather than exacerbating existing disparities or concentrating power in the hands of a few.

Many researchers may respond to this with the logistic challenges of actively enforcing guidelines on something as accessible as AI across boarders. After all, what's the use of going through years of legislative procedure just for someone to turn on their VPN and switch their location to another country. When it comes to international laws, their viability and effectiveness decreases due to jurisdictional issues, differing cultural norms, and varying levels of policing capacity among nations. Moreover, the rapid pace of technological innovation often outpaces the development of regulatory frameworks, creating a perpetual game of catch-up for policymakers. My response to that is simple: while the challenges of enforcing AI guidelines across borders are indeed daunting, they are not insurmountable. Generations before have done it before to protection people across the world from human trafficking, copyright, patent infringement etc... and so it is possible to do it for AI. As long as models continue to run on internet based cloud computing (which they will) regulation will be simple as infrastructure already exists for online illicit activity monitoring.

There's no doubt that the advancements of AI has brought a new era of efficiency, convenience, and possibilities in industries around the world at a level of accessibility never seen before. People from around the world can connect to generative AI models with simply a computer and the internet. While it is tempting to keep governments out of scientific research, a line needs to be drawn when the lives and futures are at state. This line has already been crossed multiple times due to irresponsible data labelling and discriminatory algorithms to the point where warnings no longer suffice. Governments need to reassess their laws and regulation surrounding AI and make sure that its role in the Social Contract is fulfilled and the technology's advancement brings the improvement of lives, not the persecution of it.

## Works Cited

Albaroudi, Ethan, et al. “first_page settings Order Article Reprints Open AccessReview A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring.” *MDPI*, vol. 34`, no. 2, 2024, pp. 193-201. *MDPI.com*, https://www.mdpi.com/2673-2688/5/1/19. Accessed 1 April 2024.

“Automated Vehicle Crash Sequences: Patterns and Potential.” *Accident Analysis and Prevention*, vol. 153, no. 4, 2021, pp. 39-45. *sciencedirect.com*, https://www.sciencedirect.com/science/article/abs/pii/S0001457521000488. Accessed 1 April 2024.

Christian, Brian. *The Alignment Problem: Machine Learning and Human Values*. WW Norton, 2020. Accessed 2 April 2024.

Christian, Brian. *The Most Human Human: What Artificial Intelligence Teaches Us About Being Alive*. Knopf Doubleday Publishing Group, 2012. Accessed 3 April 2024.

“Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies.” *Sci*, vol. 45, no. 19, 2023, pp. 245-256. *MDPI.com*, https://www.mdpi.com/2413-4155/6/1/3. Accessed 3 April 2024.

“Introducing Devin, the first AI software engineer.” *Cognition Labs*, 12 March 2024, https://www.cognition-labs.com/introducing-devin. Accessed 1 April 2024.

Rousseau, Jean-Jacques. *The Social Contract*. CreateSpace Independent Publishing Platform, 1762. Accessed 1 April 2024.