Towards the Responsible Development of AI

Seth Grief-Albert
Queen's University
seth.griefalbert@queensu.ca

Cyrus Fung
Queen's University
19thcf@queensu.ca

Aamiya Sidhu
Queen's University
21as226@queensu.ca

I. Introduction

Progress in Artificial Intelligence (AI) is accelerating at an unprecedented rate. The introduction of AI applications into the public domain will have profound effects on several institutions across healthcare, industry, and beyond. This necessitates reevaluating the ways in which revolutionary technology has been and continues to be developed. Social media offers a cautionary tale, that entrenched attitudes towards innovation at any cost have led to negative consequences. This attitude is incompatible with developing technology ethically. A shift in focus from rapid development to responsible development is necessary, calling for a more thoughtful approach to the creation and deployment of AI.

II. Radical Disruption

We are at a point where moving fast and breaking things has solidified itself as a driving force behind innovation. The biggest disruptors in AI are the same platforms that used this very idea to emerge as global conglomerates.

At the outset of the social media revolution, the mantra "move fast and break things" voiced by Mark Zuckerberg reflected the budding entrepreneurial attitude towards innovation [1]. This idea emerged as social media was being built as a call for fast progress and repeated iteration, at the cost of careful deployment. Social media represented a paradigm shift in communication towards a personalized network with unprecedented monetization potential. Data fuels social media. In the process of engaging with platforms, users generate information that is applicable to advertising. In 2020, over 97% of Meta's revenue was from advertising [2]. With a major market share and a highly secure profit stream, it is entirely rational for businesses like Meta to keep advertisers as clients, and make decisions with the interests of advertisers in mind.

At the same time, when technological progress moves faster than regulation and ethical guidelines, massive systems with murky purposes and emergent phenomena can arise. Without guardrails, negative consequences can easily manifest. A prominent phenomenon on social networks is the echo chamber. Echo chambers are characterized as radically exclusive social structures, with the potential to harm vulnerable online communities [3]. With little oversight and algorithms that prioritize personalization, echo chambers have established themselves throughout social media [4]. Algorithmic personalization is based on the interactions and engagement of a user, and content that someone is more likely to engage with is fed to their home screen [5]. Thus, pre-existing biases can take root, presenting a warped view of the world. During the COVID-19 pandemic, health misinformation proliferated online throughout social media [6]. Additionally, trust in social media has risen to a point where many people solely engage with it for their news [7]. The disruption of powerful technology over time paints a picture that moving fast and dismissing consequences is misaligned with the goals of beneficial and ethical application.

III. The State of AI

A. Bias in AI

From 2015 to 2021, the compound growth rate in AI-related patents was 76.9% [8]. Implementing novel developments in AI as fast as possible has produced numerous gaps, because the positive performance of AI is predicated on good data. Faulty data ravages applications in criminal law, advertising, recruiting, and computer vision [9]. It is thus necessary to evaluate the approach taken by enterprises towards data.

A black box system is characterized by inaccessible operations and explanations of outputs. In machine learning, a subset of AI, systems are created purely from data for use in algorithmic processing [10]. The output of complex systems in AI can be dictated by billions of components, potentially transcending human understanding of internal operations. Yet, the societal impact of these systems is significant. One such case is the COMPAS algorithm, which has been used to determine the likelihood of a criminal becoming a repeat offender [11]. These systems rely on historical data from eras with different laws and political views, making them highly susceptible to human bias [11]. When flawed data is applied to black box systems, a lack of transparency for developers and consumers has negative impacts.

Algorithmic bias in AI is defined as "systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over another" [12]. This bias has become one of the most pressing issues in modern AI applications. Bias is especially relevant in healthcare, where impacts can greatly affect the lives of individuals. Overarching problems with the use of AI in healthcare involve fairness, lack of context, and non-explainability. As it stands, inequities within western society and healthcare systems result in a lack of quantitative metrics of fairness. This leaves interpreters of medical evaluations responsible for reducing bias [13]. Healthcare and treatments for individuals are specific to their context which includes an individual's environment, culture, socioeconomic status, lifestyle choices and genetics

[13]. Given the number of variables, intersectional impacts, and lack of data for major population groups, AI applications in healthcare currently lack the ability to provide accurate data for underrepresented populations [13]. Because of the power and influence of AI algorithms, black box systems can leave data scientists, clinicians, and patients without knowledge of how predictions are being made. This has detrimental effects as progress accelerates and AI becomes increasingly present in healthcare [13]. The potential consequences of inequitable training data are severe. Today, several facial recognition training datasets mostly contain images of white males [9]. In tests with recent image recognition models such as CLIP, Black people could be misclassified as non-human [8].

B. *Intellectual Property*

The massive scale and black-box nature of generative models further complicates legal and ethical concerns regarding intellectual property. Models such as GPT-3, Copilot, and DALL-E 2 are trained on large amounts of data scraped from the internet, which inevitably contains copyrighted data [19]. A user could plausibly generate an output that closely resembles copyrighted data, by fine-tuning GPT-3 (a language generation model) with samples from a specific author in order to pass off its output as an original product. Models can also infringe on copyright by directly regurgitating training data in their outputs, potentially misleading users into thinking that it is a novel creation and not stolen intellectual property. For example, Copilot has been found to generate large sections of copyrighted code verbatim, leading to a class-action lawsuit against its creators for violating the rights of the millions of GitHub users who published code under an open-source license [20].

As awareness of the risk that generative models pose to intellectual property rights grows by the day, most companies filter prompts to dissuade claims that they are facilitating theft or turning a blind eye. Unfortunately, user-level interactions only represent the tip of the iceberg of copyright infringement in AI. A lack of regulation and oversight gives companies ample opportunities to exploit loopholes that allow them to violate copyright to generate profit, while also shielding themselves from accountability. One such loophole is “data laundering,” where for-profit companies strengthen their claims of fair use and profit off of copyrighted information by outsourcing the training and dataset collection process to non-profit entities [12]. Consider the text-to-image latent diffusion model Stable Diffusion by Stability AI, which has already been implicated in several copyright infringement lawsuits and is widely controversial amongst human artists [15]. One might believe that Stability AI was the creator of their flagship model, but a quick look at its GitHub repository reveals that it was trained by researchers at the Ludwig-Maximilians University of Munich using ”a generous compute donation from Stability AI” [16]. The dataset used to train the model was not collected directly by Stability AI either: instead, it came from a non-profit organization called LAION, which also received compute resources from Stability AI [17]. The degree of separation that data laundering creates protects companies in two ways. Firstly, since these models were technically created for non-profit or academic purposes, most courts would consider them as falling under fair use. Secondly, in the event of unwanted scrutiny, public attention or legal liability can be redirected from the parent corporation to the non-profit organization that created the model with its funding and guidance. This allows corporations to convert what is ostensibly academic research into a monetized product while minimizing legal risk, as seen with Stability AI’s product DreamStudio (essentially a consumer-friendly wrapper around the Stable Diffusion API). Without the appropriate regulatory oversight in place, continuing to move fast and break things in AI will only lead to unprecedented levels of intellectual property theft, harming both consumers and producers who cannot afford to keep up.

IV. Responsibly Developing AI

AI can only move as fast as the data that powers it and the regulation that guides it. Key to the responsible development of AI is transparency and accountability to stakeholders. The nature of black-box technology in machine learning applications poses a barrier to transparency and explainability. Transparency in data sourcing is an important aspect of ethical AI deployment. Data is the backbone of AI. This leads to external failure when data is non-representative, perpetuating human bias and leading to real harm [9]. Creating representative, equitable data is difficult when many companies outsource this task to third parties, thus obscuring responsibility.

Accountability in AI must be applied to every step of system development, from initial design to system monitoring. An accountable AI system encompasses four dimensions: strong governance, understanding the data, clear performance goals, and continuous monitoring [18]. The governance of a system includes ensuring the work-force is well-rounded with diverse perspective, has broad stakeholders, and strong risk-management. Accountability also requires documentation at every level. This includes technical specifications, system compliance and output, potential issues, and performance assessments [18]. Documentation must be available to stakeholders along with design and operation information. Most significantly, strong governance puts emphasis on the responsibilities of the authorities that control the deployment of the system. Accountability in a system entails a thorough understanding of the data that is used to create a model, and that understanding remains while the system is in operation. In addition, reliability and representativeness of the data must be understood to look for bias, inequities, and societal concerns from applications of the system [18]. Performance goals must be clear, well-documented, and stable from the first stages of development until the system is operational [18]. This includes performance assessments of the overall system and its individual components. The monitoring of a system entails having a set range that allows the system to “drift” and must allow for continuous questioning about the function and importance of the system [18]. Long-term monitoring assessment and changes to a system must be done to assess if the system is still working and more importantly, if the system is still required.

V. Conclusion

Charging into the future as fast as possible leaves little time for true understanding of the societal implications of AI. Although a significant body of work has already been advanced regarding AI safety, concrete implementations by governments and corporations are still lacking. This may not just be a result of inertia or competing incentives, but could also be a natural product of the black-box nature of many AI systems. This makes defining key factors such as transparency, accountability, and reliability a difficult and possibly even intractable task. Without a clear and unambiguous means of evaluating outcomes of AI applications for both government and corporate bodies, the window of responsible development of AI appears to be shrinking rapidly. Legally enforced regulatory frameworks and bodies may be necessary to ensure that AI systems adhere to the aforementioned metrics. However, these frameworks must be highly flexible to handle ambiguity, and also to keep up with the rapid evolution of AI safety going forwards. By shifting the focus from innovation at any cost to an attitude of discretion and foresight, the revolutionary potential of AI can be harnessed for the benefit of all.

References

[1] J. Liles, “Did Mark Zuckerberg say, ’move fast and break things’?,” Snopes, 29-Jul-2022. [Online]. Available: https://www.snopes.com/fact-check/move-fast-break-things-facebook-motto/. [Accessed: 12-Mar-2023].

[2] S. Dixon, “Annual advertising revenue of Meta Platforms worldwide from 2009 to 2022,” Statista, 13-Feb-2023. [Online]. Available: https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/. [Accessed: 03-Mar-2023].

[3] C. T. Nguyen, “ECHO CHAMBERS AND EPISTEMIC BUBBLES,” Episteme, vol. 17, no. 2, pp. 141–161, 2020.

[4] C. T. Nguyen (2021). How Twitter gamifies communication. In Jennifer Lackey (ed.), Applied Epistemology. Oxford University Press. pp. 410-436.

[5] Meta, “What kinds of posts will I see in Feed on Facebook?,” Facebook Help Center. [Online]. Available: https://www.facebook.com/help/166738576721085. [Accessed: 03-Mar-2023].

[6] V. Suarez-Lledo and J. Alvarez-Galvez, “Prevalence of Health Misinformation on Social Media: Systematic Review,” Journal of Medical Internet Research, vol. 23, no. 1, Jan. 2021.

[7] J. Liedke and J. Gottfried, “U.S. adults under 30 now trust information from social media almost as much as from national news outlets,” Pew Research Center, 27-Oct-2022. [Online]. Available: https://www.pewresearch.org/fact-tank/2022/10/27/u-s-adults-under-30-now-trust-information-from-social-media-almost-as-much-as-from-national-news-outlets/. [Accessed: 03-Mar-2023].

[8] D. Zhang et al., “The AI Index 2022 Annual Report,” AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022.

[9] N. Turner Lee, P. Resnick, and G. Barton, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” Brookings, 22-May-2019. [Online]. Available: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/. [Accessed: 03-Mar-2023].

[10] C. Rudin and J. Radin, “Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition,” Harvard Data Science Review, 22-Nov-2019. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8. [Accessed: 03-Mar-2023].

[11] T. Cassauwers, “Horizon Magazine,” Horizon, 01-Dec-2020. [Online]. Available: https://ec.europa.eu/research-and-innovation/en/horizon-magazine. [Accessed: 03-Mar-2023].

[12] “Research guides: Algorithm bias: Home,” Home - Algorithm Bias - Research Guides at The Florida State University, 23-Sep-2021. [Online]. Available: https://guides.lib.fsu.edu/algorithm. [Accessed: 03-Mar-2023].

[13] T. Panch, H. Mattie, and R. Atun, “Artificial Intelligence and algorithmic bias: Implications for health systems,” Journal of global health, Dec-2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6875681/. [Accessed: 03-Mar-2023].

[14] A. Baio, “AI data laundering: How academic and nonprofit researchers shield tech companies from Accountability,” Waxy.org, 30-Sep-2022. [Online]. Available: https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/. [Accessed: 01-Mar-2023].

[15] A. Martin, “Lawsuits over stability AI’s stable diffusion could threaten the future of AI-generated art,” Business Insider. [Online]. Available: https://www.businessinsider.com/stable-diffusion-lawsuit-getty-images-stablility-ai-art-future-2023-1. [Accessed: 28-Feb-2023].

[16] Stability AI, “Stability-ai/stablediffusion: High-resolution image synthesis with Latent Diffusion Models,” GitHub. [Online]. Available: https://github.com/Stability-AI/stablediffusion. [Accessed: 01-Mar-2023].

[17] R. Beaumont, “5B: A new era of open large-scale multi-modal datasets,” LAION. [Online]. Available: https://laion.ai/blog/laion-5b/. [Accessed: 01-Mar-2023].

[18] S. Sanford, “How to build accountability into your ai,” Harvard Business Review, 30-Aug-2021. [Online]. Available: https://hbr.org/2021/08/how-to-build-accountability-into-your-ai. [Accessed: 03-Mar-2023].

[19] C. Xiang, “Ai is probably using your images and it’s not easy to opt out,” VICE, 26-Sep-2022. [Online]. Available: https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out. [Accessed: 03-Mar-2023].

[20] R. Losio, “First Open Source copyright lawsuit challenges github copilot,” InfoQ, 18-Nov-2022. [Online]. Available: https://www.infoq.com/news/2022/11/lawsuit-github-copilot/. [Accessed: 03-Mar-2023].