Privacy, Prejudice, and Pixels: A Journey Through the Implications of Machine Learning

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Mitchell Hong

Spring 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Caitlin D. Wylie, Department of Engineering and Society

Introduction

Without deliberate attention to the ethical considerations involved, Machine Learning (ML) technology is rapidly integrating into every aspect of society. Improper training of machine learning models can lead to significant consequences, degrading the overall value that ML technology provides. A pertinent example of this is took place in 2017 when Idemia, a French company specializing in facial recognition technology, provided a system to the FBI that was lauded for its ability to enhance American security. The recognition technology successfully matched faces with a positive match rate of 9,999 in 10,000. In other words, the algorithm had a false match rate of only 1 in 10,000. This exceptional performance set a promising outlook for the technology's future applications. Unfortunately, this high accuracy, however, was only accurate for individuals with lighter skin tones. Later evaluations revealed a significant disparity in accuracy when the technology was applied to individuals with darker skin tones. The false match rate soared to a rate of 1 in 1,000—a rate 10 times higher than that for lighter-skinned individuals (Simonite, 2019). This discrepancy highlights the potential downfalls of poorly implemented recognition technology in high-stakes situations. Individuals with darker skin are disproportionately at risk of being falsely identified, which could lead to wrongful accusations and convictions. The stark difference in accuracy between skin tones had the potential of furthering existing biases, creating a negative feedback loop of injustice. Further, my research examines the implications of Machine Learning (ML) on data privacy, bias, and the regulatory challenges these technologies present. With careful consideration in the areas of data privacy and bias, we can develop strategies and frameworks to mitigate these risks and ensure that ML technologies are implemented in a manner that is both ethically responsible and socially beneficial. To address this complex issue, I will use the Social Construction of Technology (SCOT) and Responsible Research and Innovation (RRI) frameworks as analytical tools. SCOT posits that technology is influenced by the social, political, and economic dynamics that shape its development. In other words, technology does not determine human action, but rather, human action shapes technology (Science Direct, 2012). RRI aims to ensure the ethical acceptability, sustainability, and societal desirability of the innovation process. It challenges stakeholders to work together during the entire research and innovation process (UK Research and Innovation, 2023). While ML technology holds immense potential to contribute positively to society, it also poses significant risks, particularly if ethical considerations are not adequately addressed.

Infringements on Data Privacy

Machine Learning technologies often collect and process vast amounts of personal data without explicit consent, leading to potential privacy breaches and unauthorized data exploitation. This is particularly concerning in sensitive areas such as medical imaging, where the substantial data requirements for training ML models can lead to significant privacy issues, especially when personal health information is involved. One of the primary concerns in the application of ML is the necessity for large datasets, which often leads to the collection and use of personal data, thereby raising significant privacy concerns. A paper by Cho et al. (2016) highlights the extensive data requirements for training machine learning models in the context of medical imaging. It emphasizes the need for large, high-quality datasets to achieve high accuracy in disease diagnosis and treatment planning. However, the paper also acknowledges the challenges

in accessing such medical images due to patient privacy laws and policies, further noting the privacy issues that arise when personal health information is involved. (Cho et al., 2016).

The main issue presented is the perception of patient privacy laws and policies as mere hurdles in the ML training process. With current regulations, companies are fined for poor data collection practices (BigID, 2024). Yet, revenue generated from ML applications far outweighs the penalties and, thus, ML has proliferated throughout corporations without privacy and security as a focus. These protections were established to safeguard sensitive patient information. Viewing them as challenges instead of advantages leads to an increased risk of privacy infringements. There is a growing danger that organizations might prioritize the development of high-accuracy models over data privacy, rationalizing that the societal benefits of these models outweigh the importance of individual privacy rights. However, this stance is ethically questionable and endangers the potential well-being of many people. It is extremely dangerous to compare and measure values in this manner. Organizations should instead seek more ethical methods to collect data that comply with current privacy standards. For instance, hospitals could develop consent processes and offer incentives for patients willing to contribute their images for research purposes. Such approaches would allow the collection of a large volume of high-quality data for ML model training without infringing on data privacy.

Machine learning models can be attacked in many ways. According to Strobel and Shokri (2022), “An adversary, who can only observe the model and not the training data, can use inference algorithms to reconstruct information about the training data“ (p.2). The risk of data leakage emerges not solely from organizations that implement proper access controls and encryption but also from external entities with malicious purposes. Moreover, this situation becomes particularly concerning as malicious individuals have the ability to access and disclose confidential data. Although an adversary may not be able to reconstruct original training data completely, “learning anything about individual data records beyond the general patterns should be considered a privacy violation.” (Strobel and Shokri, 2022, p.3). The current outlook on data privacy is a matter of quantification rather than conservation. Specifically, organizations currently are prioritizing the identification of a permissible level of data privacy infringements, essentially seeking an optimal equilibrium between performance and privacy protection. However, any breach of data privacy should fundamentally be deemed unacceptable. In the Harvard Business Review, Michael Segalla and Dominique Rouziès comment, “When it comes to customer data, companies have typically been much less scrupulous. Many view it as a source of revenue and sell it to third parties or commercial address brokers” (Segalla and Rouziès, 2023, p.8). Under the methodology of Responsible Research and Innovation, the emphasis is not just on the end goals of innovation but significantly on the process through which these goals are achieved. RRI advocates for a comprehensive approach where the development of high-accuracy ML models—capable of reshaping society and driving advancements—is pursued with an ethical, inclusive, and reflective mindset (UK Research and Innovation, 2023). This framework necessitates engaging a broad spectrum of stakeholders in the innovation process, including policymakers, the public, and the end-users of technology, to ensure the outcomes are aligned with societal values and expectations. It underscores the importance of transparency, ethical integrity, and public engagement, aiming to bridge the gap between technological advancement

and societal needs. Therefore, achieving high accuracy in ML models holds true value only when the methodologies employed are consistent with these broader RRI principles.

Ensuring data privacy during the ML training process extends beyond safe and ethical data collection methods. The technical strategies employed during model training also play a crucial role in safeguarding data. It is essential to focus on mitigating potential data leakage, or the ability to extract training data, from models. Further, various techniques are currently being developed to train ML models securely. Examples include differential privacy, which adds noise to the data to prevent identification of individuals, and federated learning, which trains algorithms across multiple decentralized devices without exchanging data samples (BigID, 2024). The adoption, however, of these techniques by organizations has been relatively slow.

Federated learning and differential privacy techniques are being employed to address the ethical privacy issues in the ML training process. Problems of these techniques are addressed in the paper by Huang et al. (2020) where they state, “data samples distributed across different platforms are not independently and homogeneously distributed. When dealing with non-IID (independent and identically distributed) data, the training complexity of the federated learning model may be greatly increased.” (p.2). This complexity is further increased by differential privacy, where intentional noise is added to the data to defend against inference attacks. Both of these methods greatly increase the complexity of the data, driving accuracy down in the process. In sectors such as medicine and autonomous vehicles, decreases in accuracy by even a small number of percentage points can lead to unacceptable ramifications. This leads to a misalignment between organizational goals, which aim to develop high-value machine learning models for societal benefit, and individual expectations, which seek to gain high-value services from these organizations without compromising personal privacy or interests. Further, an iterative process of innovation, guided by RRI is necessary. This process would seek to balance organizational goals with individual rights and societal expectations, fostering an ecosystem where technological advancements in ML contribute positively to societal challenges without sidelining ethical considerations. By intensifying research initiatives within this framework, the aim is to cultivate reliable, high-precision machine learning models that uphold rigorous data privacy standards.

These papers were all published on arXiv, a free distribution service and an open-access archive for scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics. It allows research papers to be shared without being published in a peer-reviewed journal. While arXiv is not a peer-reviewed journal itself, arXiv employs a moderation process to ensure submissions are relevant and of a certain quality. The oversight by Cornell University enhances arXiv's reputation within the scientific community. The arguments presented in these papers align well with the prevailing discourse on this subject. Additionally, these papers have not received any reprimand or critiquing from the scientific literature community. Based on this analysis, I consider these articles to be both credible and pertinent.

Perpetuation of Bias

ML algorithms, trained on historical data, can perpetuate existing societal biases, leading to unfair and discriminatory outcomes in areas like employment, law enforcement, and credit scoring (Agarwal et al., 2023). The efficacy of a machine learning model is constrained by the quality of its training data. The adoption of ethical and responsible data collection practices has been a relatively recent development. Consequently, data sets have often inadvertently captured external trends and patterns that may not be directly relevant to the intended analysis. Consider the following thought experiment: a manufacturer of fighter jets gathers data exclusively from aircraft that have returned from combat. Following an analysis, the decision is made to reinforce the sections of the aircraft that show the highest concentration of bullet damage. However, an external reviewer posits a critical question: is it truly advantageous to fortify the parts of the jet that sustained damage yet still managed to return safely? It could be argued that reinforcing the areas that, if damaged, would prevent the aircraft from returning at all would be more effective. The main obstacle to this counterargument lies in the absence of data on the aircraft that did not make it back. This exemplifies Survivorship Bias, a prevalent issue in data collection processes (Katopol, 2017). In the context of machine learning, a model cannot fully serve the interests of all stakeholders if the data it trains on does not comprehensively and equitably represent the entire spectrum of invested parties.

Ensuring algorithmic fairness in machine learning is crucial to prevent the perpetuation of biases. Even if data is collected in a manner that's fair and representative of all stakeholders, an unfair algorithm may produce a model favoring certain features in the data. Consider this situation: in sectors such as the medical industry, the detection of threatening conditions is paramount. A hospital collects a vast amount of data on patients with the goal of constructing a model to predict the likelihood of an individual having a specific disease (Aung et al., 2021). The data was collected on patients representing a plethora of different backgrounds, ethnicities, and health conditions. One might assume that due to the robust data gathering methods, the construction of a fair, effective model is a likely outcome. This assumption, however, is significantly misguided. The first misconception in this order of logic is that the model will successfully predict real-world outcomes even if designed without careful consideration of its underlying assumptions and variables. In the highly lauded and respected journal within the scientific community, "Nature Machine Intelligence" , Mhasawade et al. (2021) illustrates the complexity of model design by stating, "although accounting for factors such as 'race' may be important in specific analyses, it is often unknown what the comprising factors of such social constructs are, how they interact and how to model them" (p.5). Every paper published in this journal undergoes a thorough peer review process, ensuring that only research of the highest quality and significance is shared with the wider community. Moreover, while the data gathered may encompass a broad spectrum of individuals, the underlying framework defining what the data truly represents is often ambiguous. For instance, race is the generalization and loose grouping of individuals. This, however, undermines the biological variation between individuals and threatens equity in representation. With this mind, the model being constructed should focus less on race and more on the relationships between other features in the data. For societal issues such as poverty, inequality, and racism, race can be more loosely coupled from genetic variation, but in a medical sense, race must be looked at under a more detailed lens. The second misconception is the belief that correlation implies causation. Frequently, features in the data may show strong correlations

with the outcome the model aims to predict. However, this does not guarantee that such features are causally connected to the predicted outcome. Moreover, prior to the construction of the model, it's crucial to establish a threshold for significance. By doing so, statistically significant relationships and those that might occur by chance can be differentiated between, helping to focus on the most impactful predictors.

Machine learning algorithms can inadvertently reinforce existing disparities, affecting marginalized populations. This is especially problematic in sectors where ML technology is being employed to ensure individuals have equal access to opportunities. Published by the American Journal of Public Health (AJPH), a highly reputable and trustworthy journal where papers are critically assessed by independent experts in their peer reviewing process, Samorani and Blount (2020) examines the use of ML in medical appointment scheduling, illustrating how these algorithms might unintentionally discriminate against certain groups, particularly those of lower socio-economic status or with less reliable transportation. They outline the overbooking process clinics use to maximize efficiency – ML algorithms purposely overbook individuals with the lowest probability of showing up. In practice, these models effect primary African Americans. Samorani and Blount (2020) state, “it is well known that lower show-up probabilities are correlated with factors typically associated with less advantaged socioeconomic status: limited transportation, lack of health insurance, and inconsistent employment, to name a few” (p.1). The original intention of implementing this ML model was to maximize efficiency in clinics, attempting to provide the most value possible for people living in these areas. The outcome, however, ended up impacting already marginalized groups. Without correction in these systems, a positive feedback loop will occur over time. In more detail, patients with less advantaged socioeconomic status will find difficulty in receiving the treatment they need. This in turn has potential implications on employment, further negatively impacting their socioeconomic status. Before ML implementation initiatives begin, significant effort must be put into the data collection process. Although this specific case study only highlights issues in the health care sector, these issues are found in many other sectors as well, including finance, education, and government. Without a change in the way ML algorithms are handled, the consequences will far outweigh the value brought by ML.

Under the Social Construction of Technology framework, it is argued that technology does not shape human behavior, and vice versa. Rather, it is the interaction between different social groups and technology that shapes the design of future technologies (Science Direct, 2012). These groups could include developers, funders, users, and those affected by algorithmic decisions. With this, there are a few main possibilities as to why machine learning algorithms are perpetuating biases and reinforcing already existing disparities. One possibility is that society designs these systems with the goal of perpetuating biases. Another possibility is that there is a major disconnect between the implementation and the intended design of these systems. I, however, argue that the perpetuation of bias and reinforcement of already existing disparities simply arises from the immense complexity of values, priorities, and knowledge-banks of connected groups. Each of these factors build upon one another, creating a web of complexity. In order to make meaning of the madness, we must place more focus on the groups – stakeholders

and technologies — that are directly and indirectly involved. Under SCOT, these factors must be analyzed to properly address current issues and provide fair and equitable ML to all of society.

Challenges of Regulations

The rapid advancement and integration of ML technology outpaced the development of regulatory frameworks, leading to a lag in addressing ethical concerns and societal impacts. As with other disruptive technologies, regulatory bodies encounter the Collingridge dilemma. They need to decide whether to regulate a technology during its nascent phase or to hold off until it becomes more established. Opting for early regulation presents a challenge because the novelty of the technology means its full implications are not yet understood by society. On the other hand, waiting too long to introduce regulations could lead to them being less impactful -- the technology would have already evolved without constraints (Kudina and Verbeek, 2018). Society has opted for the latter, allowing ML technology to develop fully unconstrained. There are no established standards for data quality, algorithmic benchmarks, or a governing authority. Businesses have the liberty to pursue their own machine learning projects initiatives, utilizing data they've gathered and algorithms they've selected independently. Transparency regarding the accuracy and efficiency of their models has not been a priority, as there's been no imperative to do so. The allure of ML technology as a buzzword garners interest, irrespective of the actual success of its application.

Developments in machine learning have outpaced regulatory measures, leading to significant gaps in addressing emerging ethical and societal concerns. Instead of debating whether ML technology should be introduced to society, Vesnic-Alujevic et al. (2020) states that we should "rather understand 'how' to live with these technologies, what is their current meaning in specific social, political or cultural, individual and collective settings, and how we can inform about their development in the near future" (p.1). Published in the Telecommunications Policy journal, an international, reputable journal that undergoes a thorough peer reviewing process, Vesnic-Alujevic et al. present a credible and insightful perspective. Machine learning development must utilize a holistic approach, prioritizing macro impacts over micro focuses. Concentrating on a single aspect, such as the economy, may lead to significant changes within that domain. In spite of this, it's crucial to consider the repercussions on other sectors. Employing ML technology to transform the economic landscape could inadvertently reshape the cultural and social climate, potentially leading to job losses, an increased wealth gap, and further marginalization of vulnerable groups. Policy makers have approached regulation with a micro mindset, failing to address the complexity and interconnectedness of ML technology. The issue extends beyond mere data privacy concerns. Regulators are now tasked with addressing a broader array of challenges, including unemployment, equity, bias, safety, and the impact on human behavior. As ML technology advances, regulatory frameworks must evolve correspondingly to prevent the marginalization of specific groups.

Conclusion

The critical issues of data privacy and bias substantiate that while ML technology offers significant benefits in many sectors of society, it also poses substantial risks if ethical

considerations are not properly addressed. The implication of my claim suggests a pressing demand for enhanced privacy protection, fair algorithmic practices, and responsible consumer engagement. The main limitation is that the constantly evolving nature of ML technologies further increases the complexities involved in regulatory practices. This in turn creates an environment where ethical standards are dynamic and nebulous. Fortunately, it is humans who are the drivers of change and progress. The outlook that technology dictates the trajectory of society is both outdated and incorrect. With this in mind, as society continues to develop ethical and moral guidelines, technologies will be designed to align with these advancements. The cornerstone of future success lies in intentionality. Given the unanimous standards of moral and ethical virtue already acknowledged by society, integration is the next logical step. Historically, mathematicians and scientists understood that working with biased data would result in biased outcomes. Leveraging this knowledge, they avoided working with biased data to ensure that the results produced were representative and useful. While ML technology is certainly disruptive, it is not a fix-all solution. Future research topics that need to be addressed are determining which algorithms are considered fair, how to better model real-world concepts in machine learning models, and simulating second, third, and fourth order effects of regulations. Approaching ML development with the same precision a scientist would apply to a molar conversion – using sterile equipment, exact measurements, etc. -- will lead to the creation of models that are equitable, fair, and effective.

References

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702. https://doi.org/10.1016/j.hlpt.2022.100702

Aung, Y. Y., Wong, D. C., & Ting, D. S. (2021). The promise of Artificial Intelligence: A review of the opportunities and challenges of artificial intelligence in Healthcare. *British Medical Bulletin*, *139*(1), 4–15. https://doi.org/10.1093/bmb/ldab016

BigID. (2024, April 23). *Navigating AI data privacy: Current hurdles, future paths*. https://bigid.com/blog/navigating-ai-privacy/

Cho, J., Lee, K., Shin, E., Choy, G., & Do, S. (2016, January 7). *How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?* arXiv.org. Retrieved March 6, 2023, from https://arxiv.org/abs/1511.06348

Ferm, L.-E. C., Thaichon, P., & Quach, S. (2023, August 30). *Routledge . AI and its implications for Data Privacy*. https://www.routledge.com/blog/article/ai-and-its-implications-for-data-privacy

Güngör, H. (2020). Creating value with Artificial Intelligence: A multi-stakeholder perspective. *Journal of Creating Value*, *6*(1), 72–85. https://doi.org/10.1177/2394964320921071

Huang, W., Li, T., Wang, D., Du, S., & Zhang, J. (2020, December 18). *Fairness and accuracy in federated learning*. arXiv.org. https://arxiv.org/abs/2012.10069

Kaplan, J. (2016). Artificial Intelligence - What Everyone Needs to Know. Oxford University Press. https://doi.org/10.1093/wentk/9780190602383.001.0001

Katopol, P. (2017). Maybe best practices aren't: how survivorship bias skews information gathering and decision-making. *Library Leadership & Management*, *32*(1). https://doi.org/10.5860/llm.v32i1.7287

Kudina, O., & Verbeek, P.-P. (2018). Ethics from within: Google Glass, the collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values*, *44*(2), 291–314. https://doi.org/10.1177/0162243918793711

Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine Learning and Algorithmic Fairness in public and Population Health. *Nature Machine Intelligence*, *3*(8), 659–666. https://doi.org/10.1038/s42256-021-00373-4

Segalla , M., & Rouziès, D. (2023, June 13). *The ethics of Managing People's Data*. Harvard Business Review. https://hbr.org/2023/07/the-ethics-of-managing-peoples-data

Samorani, M., & Blount, L. G. (2020). Machine learning and medical appointment scheduling: Creating and perpetuating inequalities in access to health care. *American Journal of Public Health*, *110*(4), 440–441. https://doi.org/10.2105/ajph.2020.305570

Science Direct. (2012). *Social Construction of Technology*. Social Construction of Technology - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/social-sciences/social-construction-of-technology

Simonite, T. (2019, July 22). *The best algorithms still struggle to recognize Black Faces*. Wired. https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/

Strobel, M., & Shokri, R. (2022). Data Privacy and trustworthy machine learning. *IEEE Security & Privacy*, *20*(5), 44–49. https://doi.org/10.1109/msec.2022.3178187

UK Research and Innovation. (2023, March 16). *Framework for Responsible Research and Innovation*. UKRI. https://www.ukri.org/who-we-are/epsrc/our-policies-and-standards/framework-for-responsible-innovation/

Vesnic-Alujevic, L., Nascimento, S., & Pólvora, A. (2020). Societal and ethical impacts of Artificial Intelligence: Critical Notes on european policy frameworks. *Telecommunications Policy*, *44*(6), 101961. https://doi.org/10.1016/j.telpol.2020.101961