Proceedings of CONF-CDS 2025 Symposium: Data Visualization Methods for Evaluation
DOI: 10.54254/2755-2721/2025.PO25142

Mandating Transparency: A Regulatory Framework for Algorithmic Bias and Opacity

Kaichen Zhang

Beijing National Day School, Beijing, China
kaichenzhang76@email.com

Abstract. Algorithmic bias and opacity pose threats to the global population by damaging social equity, as demonstrated by racial disparities in facial recognition accuracy and healthcare prediction algorithms. This is mainly because most algorithms are black boxes---few know how they operate; sometimes this is intentional to maximize company profits. While ethical principles lack enforceability, current regulatory approaches like the EU AI Act focus primarily on risk and privacy but fail to identify and address biases directly. A disclosure-based regulatory framework adapted from the US FDA's Total Product Lifecycle (TPLC), originally designed for medical algorithms, is a favorable alternative. Extended TPLC has already been proposed, which lists out potential biases throughout the entire implementation of medical algorithms. As a successfully applied visual diagnostic AI, the IDx-DR demonstrates the effectiveness of companies referencing the TPLC to identify, address, and publicly disclose their handling of biases throughout an algorithm's lifecycle. This shows the potential of this framework in resolving algorithmic bias and corporate secrecy ("black boxes") in all fields. However, biases and disclosure requirements of companies need to be generalized for wide applications. The extended TPLC should thus be simplified, involving only three stages: design, development, and deployment. Its effectiveness is demonstrated through three examples: flawed teacher evaluation algorithms (design), specialized legal AI development (development), and biased healthcare risk score monitoring (deployment). While this framework promotes algorithm transparency and accountability without stifling innovation, its effectiveness can be enhanced if governments mandate disclosure accordingly and coordinate globally to prevent corporate regulatory arbitrage.

Keywords: Algorithmic Bias, Opacity, Total Product Life Cycle (TPLC), Regulatory Transparency

1. Introduction

Algorithms, including but not limited to forms of Artificial Intelligence, are already being deployed to govern our lives [1-3]. That means that superficially innocuous errors, presenting ‘mere’ technical challenges to be resolved by tweaking the algorithm or improving the dataset, are important. In the “Gender Shades Study,” it emerged that commercial facial analysis algorithms were more accurate for people with lighter skin, and particularly less accurate for darker-skinned women [4]. Security

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.PO25142

agencies have already implemented facial recognition software to discern suspects, making the results of any error potentially fatal [5]. Algorithms are biased when they favor certain conditions, which might include gender or race, thus leading to potentially harmful discrimination and limiting social equity. Yet, these can be difficult for users to recognize, and, in the case of AI, challenging even for the developers to understand.

Users must identify bias through self-validation, and may encounter algorithms (e.g., a product recommendation algorithm) about which they have been provided with no data. In this case, they come up against corporate opacity. Zuboff [6] uses the term "Surveillance Capitalism" to describe how technology companies deliberately leave consumers ignorant to "claim human experience as free raw material for hidden commercial practices of extraction, prediction, and sales", suggesting, in line with Pasquale, that internal systems of algorithms remain "black boxes" to the wide range of users. Companies gain immense profit from using people's private information to train algorithms without their consent or understanding. This disparity in acquired information is deliberate and often hidden behind the claim that they are business secrets [3].

Often, the operation performed by the algorithm has no easy parallel in human thought, given the scale of the multivariate data analysed, as exemplified by Machine Learning (ML) algorithms. For instance, Large Language Models (LLMs) recognize syntax patterns from an enormous dataset to generate text sequences [7]. Any specific output of such an algorithm is mostly unexplainable even to the AI developers. In what follows, I argue for a framework that asks companies to explicitly test for, identify, and resolve bias while ensuring transparency from design to deployment.

2. Principles or regulation

Can opacity be dealt with using existing approaches? So far, governments have been wary to act, while researchers and companies have proposed their principles. The Montreal Declaration, an initiative of researchers at the University of Montreal with signatories from across research and business, offers an ethical framework for developing and deploying AI. It suggests key principles such as enhancing human well-being, ensuring equity, and respecting autonomy [8]. However, the effectiveness of principles-based approaches mainly relies on the goodwill of researchers and technology companies. For instance, Google fired its AI ethics unit founder and co-head, Margaret Mitchel, ostensibly for data breaches; she was a fierce critic of Google and pinpointed multiple ethical concerns in AI development [9].

An alternative is for governments to regulate. The EU AI ACT, touted as "the first comprehensive AI regulation in the world" [10], offers a certification framework centered around the different levels of risk posed by uses for AI. Several practices, including "purposefully manipulative or deceptive techniques" and "exploiting vulnerabilities of people", are explicitly prohibited. Regulation of this kind is minimally burdensome for companies because it shapes the direction of AI development through disclosure rather than extensive prohibitions or incentives..

3. The FDA TPLC model

Governments should adopt a certification regulatory system through referencing revised versions of the Total Product Lifecycle approach (TPLC), an ethical framework to guide oversight of medical devices used by the US FDA. This has already been partially adapted by Abràmoff et al. [11] to center equity considerations, pinpointing specific biases of which developers should be aware at six stages of AI implementation (shown in Figure 1). The information disclosed by companies is published in a public database.

DOI: 10.54254/2755-2721/2025.PO25142

In comparison to the EU AI ACT, this has the advantage of explicitly addressing the issue of bias (the AI ACT is framed around concerns of privacy and automated decision making). Despite emerging from the specific questions relevant to medical devices, it still highlights concerns relevant to all forms of algorithms. Operating through disclosure and oversight, it can be adapted by different regulatory agencies to accommodate their needs. By contrast, the AI ACT only separates AI into specific and general-purpose AI. Furthermore, explicitly mandating these bias disclosures in all sectors challenges the "black box" of algorithm development described by Pasquale and Zuboff.

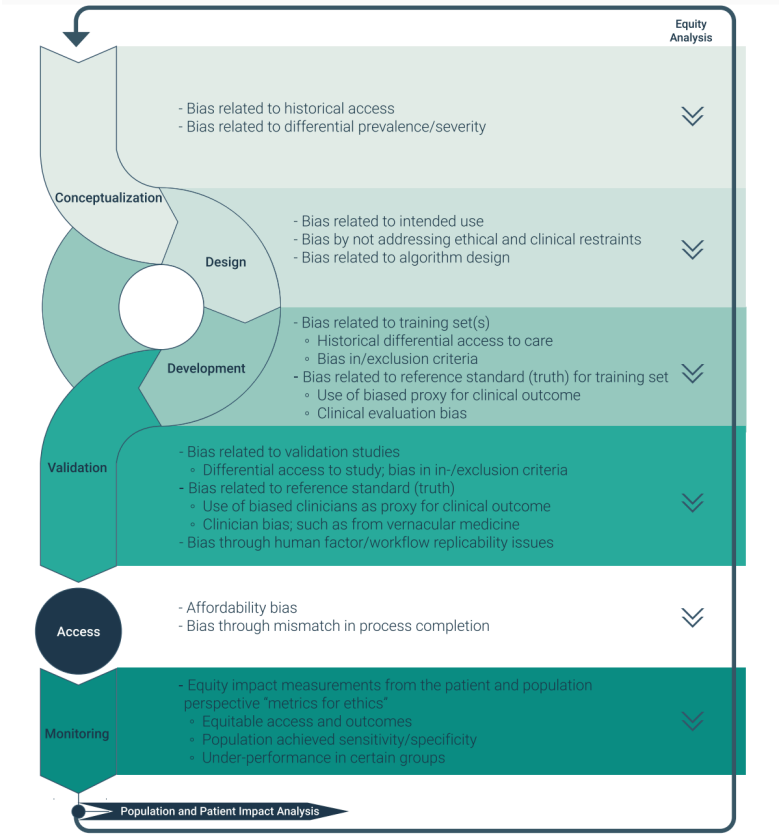


Figure 1. The extended Total Product Life Cycle (TPLC) [11]

This extended TPLC has already been implemented in the development of IDx-DR, a fully automated AI to detect diabetic retinopathy (DR) [12]. One can see how the biases suggested by the framework have been addressed in its design. For instance, IDx-DR excels at providing indications and explanations for its conclusion, allowing clinicians to validate its diagnosis, demonstrating the developers' reflection on intended use. It utilizes various DR indicators as biomarkers for the AI to process before reaching its conclusion, improving result accuracy, and addressing possible bias related to the reference standard (truth).

The developers also designed multiple metrics to continuously monitor the stability of the AI, including sensitivity, specificity, and diagnosability. These metrics measure the accuracy and consistency across different subgroups of diagnosed results, thus establishing a comprehensive evaluation system for validation. This validation test revealed that IDx-DR was reliable in generating accurate diagnoses without racial bias, which led the FDA to authorize it for use in healthcare settings [13].

DOI: 10.54254/2755-2721/2025.PO25142

4. Adapting the TPLC

Often, the operation performed by the algorithm has no easy parallel in human thought, given the scale of the multivariate data analyzed, as exemplified by Machine Learning (ML) algorithms. For instance, Large Language Models (LLMs) recognize syntax patterns from an enormous dataset to generate text sequences [14]. Any specific output of such an algorithm is mostly unexplainable even to the AI developers. In what follows, I argue for a framework that asks companies to explicitly test for, identify, and resolve bias while ensuring transparency from design to deployment.

The extended TPLC is limited for implementation in sectors other than healthcare because multiple biases are specific to healthcare settings, such as bias from not addressing clinical constraints and bias in clinician evaluation. These biases do have parallels; for instance, studies on self-driving cars have shown that drivers frequently struggle to remain sufficiently alert to intervene in potential emergencies, demonstrating consistent bias in their evaluation of the capacities of the car framework (a bias in driver evaluation) [15]. Therefore, the TPLC should be adapted to account for variations. The exact requirements should be flexible and need not be onerous when applied to fields where human lives aren't at stake and the negative consequences of algorithmic bias and opacity are less severe. For instance, very few of these biases might be relevant to algorithms developed to solve user-inputted mathematical problems, such as Wolfram Alpha.

I argue that regulation should focus on design, development, and deployment, as these three stages simplify the TPLC for application to a broad range of sectors. Companies should disclose how they have responded to each potential source of bias and publish the results publicly. While this is part of the TPLC regulatory framework suggested by the FDA, the effect of applying this to other sectors, including the bank fraud detection algorithms that Eubanks criticizes, would be to systematically challenge corporate opacity.

Specific implementations would vary to ensure that governments do not intervene excessively in sectors where competition is effective in improving products. Regulators should be aware that the purpose of this framework is not to restrict technology development, but to support innovation while addressing ethical concerns. In many sectors, companies may publish documents to show their achievements in regulation by presenting how they minimized specific biases in each of the three stages. Government agencies may then certify or authorize the algorithms according to the documents and conduct further monitoring. This type of regulation can resolve biases and opacity for various algorithms in different fields, as demonstrated by the following examples.

4.1. Design

In 2009, Washington, DC, introduced IMPACT, a new teacher evaluation system [16]. 50% of the evaluation was generated by an algorithm that calculates a 'value-add' score for teachers, based on the increase in students' grades from year to year, with some weighting for individual characteristics. The mayor hired an education reformer who decided to fire teachers "in the bottom 2 percent", as they were considered low-performing teachers who would limit students' learning efficiency. Yet, if the developers had referenced the adaptation of the TPLC framework, they would have realized the bias in algorithmic design, as external factors could significantly influence the value-adding score. For instance, a class with top students is less likely to experience a significant increase in average grades compared to a class with students who initially have lower grades. The sample size was also likely to be insufficient. Perhaps they would have decided to use more metrics rather than a single 'value-add' score. They might have considered the benefit of greater transparency in how the score

DOI: 10.54254/2755-2721/2025.PO25142

was generated so that teachers could understand and offer critique, improving the accuracy of the product.

4.2. Development

The regulatory framework also applies to more complex algorithms, as exemplified in the field of legal AI. Östling et al. point out that legal language is distinct from natural language, with terms being used both exceedingly precisely and in ways that do not match their use in other contexts [17]. Furthermore, the UK, where the researchers are based, operates a case law system. This means that in addition to the laws issued by the government, lawyers must consider the previous decisions of judges, which can date back hundreds of years [18]. These features pose unique challenges to addressing bias in the reference standard, in this case, truth. Östling et al. explicitly considered this kind of bias in designing a solution, as they would have been required to do under the TPLC framework. They created the Cambridge Law Corpus (CLC), consisting of over 250,000 UK court cases, and chose to fine-tune the dataset through annotation rather than developing a new model [17]. The result was a highly accurate legal AI.

4.3. Deployment

Continued monitoring throughout AI deployment to identify biases related to equity and underperformance in certain populations is essential, ensuring that undiscovered biases do not become permanent product features. Obermeyer et al. conducted tests on a widely applied computer algorithm that generated risk scores to indicate which patients needed extra help in managing complex health conditions [19]. They discovered that “at the same level of algorithm-predicted risk, Blacks [had] significantly more illness burden than Whites”, indicating the algorithm was biased and undervalued the sickness of Black patients. They further revealed that one of the proxies for patients’ health conditions was predictions of future healthcare costs. This proxy neglected historical access disparities, as Black patients generally received less medical care throughout previous decades. Therefore, the predicted costs for their health care were commonly lower. Had the output of the AI not been monitored, the center would have simply continued to favor providing additional support to white patients. As a result, the algorithm was redesigned to use alternative proxies instead of predicted healthcare costs.

5. Conclusion

Issues with algorithmic bias and opacity require serious intervention from governments. I have suggested one flexible framework, whereby companies publicly disclose how they have responded to different potential issues with bias, which would reduce and potentially challenge corporate opacity. Imagine, for instance, understanding why Amazon keeps recommending you useless eye cooling patches! Or how it systematically favors its own products, acting as both a marketplace and a seller. Having this disclosed publicly might support the antitrust effort (Stoller). However, the challenge with technology is that if only a single nation or bloc adopts these solutions, then corporations will flee those jurisdictions in favor of more lax regulators. Even the threat of prohibiting their entry into the market is limited - discouraging AI development might potentially impair the economy as a whole, leaving entire sectors less capable of outperforming foreign competition. We therefore need global coordination on AI, including a serious discussion of how we

DOI: 10.54254/2755-2721/2025.PO25142

can ensure that development remains human-centered, enhancing human capacities rather than replacing them.

References

[1] Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

[2] O'neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Penguin Books

[3] Pasquale, F. (2015). Black box society : The secret algorithms that control money and information. Harvard University Press.

[4] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1–15. https: //proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[5] Frenkel, S., & Odenheimer, N. (2025, April 25). How the war in gaza drove israel's A.I. experiments. The New York Times. https: //www.nytimes.com/2025/04/25/technology/israel-gaza-ai.html

[6] Zuboff, S. (2019). The age of surveillance capitalism. Profile Books. (Original work published 2018)

[7] Murgia, M. (2023, September 9). Generative AI exists because of the transformer. Subs.ft.com. https: //ig.ft.com/generative-ai/

[8] Montréal declaration on responsible AI. (n.d.). Déclaration de Montréal IA Responsable. https: //montrealdeclaration-responsibleai.com

[9] Margaret mitchell: Google fires AI ethics founder. (2021, February 20). BBC News. https: //www.bbc.com/news/technology-56135817

[10] Fernhout, F., & Duquin, T. (2024, February 13). The EU artificial intelligence act: Our 16 key takeaways | stibbe. Www.stibbe.com. https: //www.stibbe.com/publications-and-insights/the-eu-artificial-intelligence-act-our-16-key-takeaways

[11] Abràmoff, M. D., Tarver, M. E., Loyo-Berrios, N., Trujillo, S., Char, D., Obermeyer, Z., Eydelman, M. B., & Maisel, W. H. (2023). Considerations for addressing bias in artificial intelligence for health equity. Npj Digital Medicine, 6(1), 1–7. https: //doi.org/10.1038/s41746-023-00913-9

[12] Abràmoff, M. D., Tobey, D., & Char, D. S. (2020). Lessons learned about autonomous AI: Finding a safe, efficacious, and ethical path through the development process. American Journal of Ophthalmology, 214, 134–142. https: //doi.org/10.1016/j.ajo.2020.02.022

[13] Abràmoff, M. D., Whitestone, N., Patnaik, J. L., Rich, E., Ahmed, M., Husain, L., Hassan, M. Y., Tanjil, M. S. H., Weitzman, D., Dai, T., Wagner, B. D., Cherwek, D. H., Congdon, N., & Islam, K. (2023). Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. Npj Digital Medicine, 6(1), 1–8. https: //doi.org/10.1038/s41746-023-00931-7

[14] Hirsch, M., Diederichs, F., Widlroither, H., Graf, R., & Bischoff, S. (2020). Sleep and take-over in automated driving. International Journal of Transportation Science and Technology, 9(1), 42–51. https: //doi.org/10.1016/j.ijtst.2019.09.003

[15] Li, X., Chen, Z., Zhang, J. M., Sarro, F., Zhang, Y., & Liu, X. (2024). Bias behind the wheel: Fairness testing of autonomous driving systems. ACM Transactions on Software Engineering and Methodology, 34(3). https: //doi.org/10.1145/3702989

[16] Östling, A., Sargeant, H., Xie, H., Bull, L., Terenin, A., Jonsson, L., Magnusson, M., & Steffek, F. (2023). The cambridge law corpus: A dataset for legal AI research. ArXiv.org. https: //arxiv.org/abs/2309.12269

[17] Odering, J. (2024). Library guides: United kingdom law: Case law. Unimelb.libguides.com. https: //unimelb.libguides.com/c.php?g=925964& p=6688470

[18] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. https: //www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ziad_obermeyer.pdf

[19] Stoller, M. (2020). Goliath the 100-year war between monopoly power and democracy. New York Simon and Schuster Paperbacks October.li