Mitigating Algorithmic Political Bias Through Ethical AI Policy and Governance

Christopher Fornesa

Faculty of Computing and Data Sciences, Boston University

DX 701: Responsible and Ethical Data Science and AI

Dr. Seth Villegas

August 10, 2025

Problem: How Algorithmic Political Bias Encourages Algorithmic Discrimination

A significant ethical issue in Media and Entertainment is the problem of algorithmic political bias, resulting in increasing political polarization, partially due to the virtuality and neutrality fallacies. In *Ethics and Technology*, Tavani invoked James Moor's explanation of the virtuality fallacy on the premises that "X exists in cyberspace" and that "Cyberspace is virtual, thereby, "X (or the effect of X) is not real" (p. 83). Additionally, Bezzubova explained that "digital depersonalization is closely linked to a lack of 'wholeness of perception' and 'wholeness of relatedness'" (2018). Despite the real impacts of online actions, online political polarization is perceived as its serious impacts aren't viewed as "whole," even though they are real. Many internet users also assume that algorithms are objective. Weerts, et al. coined this the neutrality fallacy, built on the "assumption that the data that is used to train the model is 'neutral' towards a given social reality" (2024). In such ways, the virtuality and neutrality fallacies enhance algorithmic political bias due to current public perception and societal norms.

Peters also outlined that, "while AI systems enjoy an aura of objectivity and accuracy… they can show algorithmic bias", which is when algorithms operate in favor or against certain social groups and is often implicit. Peters also mentioned how interest in the political preferences of individual users has resulted in algorithms that regularly track users' political preferences as part of processes to build user profiles based on personality preferences. As algorithms can fulfill tasks efficiently without fail, algorithmic political bias has an inherently maximized impact compared to offline political polarization. And, without real pressures "to temper disapproval of political opponents" (2022), algorithmic political bias has become the norm. Thus, curbing algorithmic political bias requires consequences to ensure accountability for harm as, otherwise,

entities may prioritize their own motives over the public good. Whether due to external pressure, individuals can sacrifice their own sense of duty, character or virtue to capitulate to appeal to authority, with entities' interests as justification. As algorithmic political bias is often embedded and implicit, active regulatory and consequentialist measures are necessary.

Case Studies: Societal Harm Caused by Algorithmic Political Bias and Polarization

There have been several cases of algorithmic political harm, such as "epistemic bubbles" which isolate users into in-groups built on shared political interests using clustering techniques by social media companies. These bubbles rewire the personal attitudes of online social media users "to reflect the opinions of the people they're linked to" and may radicalize these users. Kelly also stated how economic anxiety increases polarization and leads to identity-based conflicts (2021). Miles Klee also outlined how XAi's chatbot, Grok, brought up the myth of "white genocide" in South Africa and Holocaust denial after being prompted by users for unrelated topics "due to an unauthorized change to my programming" (2025). This instance highlights concerns that bad actors can manipulate algorithms with bias against marginalized groups. Veluru noted when a deep fake of former President Barack Obama was circulated by a team supervised by Dr. Suwajanakorn at the University of Washington (p. 3, 2023). He also specified that any usage of deep fakes pose "numerous ethical and security risks." (p. 3, 2023). This instance showcases why consequentialist policies must be implemented to mitigate harm.

Case Study: The 2025 Minnesota Democratic Party Assassination Attempts

On July 14, 2025, Vance Boelter killed Minnesota State Representative Melissa Hortman and her husband, and injured Minnesota State Senator John Hofmann.

Stakeholder 1: Direct victims and unharmed Democratic Party Politicians.

In addition to assassinating Representative Melissa Hortman and her husband and attempting to assassinate Senator John Hofmann, authorities found a hit list in Boelter's vehicle that named Hortman, Hoffman, and other Minnesota Democratic Party politicians, including "Gov. Tim Walz, U.S. Rep. Ilhan Omar, U.S. and state Attorney General Keith Ellison."

Stakeholder 2: The Assassin, Vance Luther Boelter.

Boelter has been described as a father of five who has worked in the food, funeral, and security industries with extreme political views regarding the LGBTQ+ community and abortion rights. Those familiar with Boelter also described him as an ardent supporter of President Donald Trump. Boelter has also been noted for his use of social media to spread these ideologies over the years before committing these crimes (Shapiro & Margolin, 2025).

Stakeholder 3: Rightwing Social Media Influencers - Senator Mike Lee, Mike Cernovich, and Laura Loomer.

Leingang and Gedeon attributed the misinformation campaigns waged by what they called the "rightwing media ecosystem" during the aftermath of tragedies, whether these claims have been substantiated by authorities. Specifically, Senator Mike Lee, in addition to other conservative media pundits, such as Mike Cernovich and Laura Loomer, spread misinformation about Boelter's perceived ideological stances in the immediate aftermath of these assassinations, even before official reports by authorities had been released. Whereas Lee claimed that Boelter was a follower of what he called "Marxism", Cernovich and Loomer falsely stated that Governor Walz was the orchestrator of these crimes, despite his mention as a target in Boelter's list (2025).

Policy Advocacy & Feasibility: Explainable AI Policy to Mitigate Algorithmic Political Bias

The Case for Algorithmic Transparency to Combat Algorithmic Discrimination.

To guide the pursuit of algorithmic fairness, I posit a policy to mandate algorithmic transparency, fit with mandatory oversight and meaningful consequence measures, to mandate explainability in AI. This will increase algorithmic transparency as a necessary first step to solving unethical AI issues (including political polarization). But even algorithms used to detect bias can have embedded bias. As such, entities may not effectively self-regulate towards explainable AI due to the implicit bias, the profit motive, or political motives.

Wang, et al. mentioned that "algorithmic systems can engage in discrimination through disparate impact, even if they do not explicitly use protected characteristics or proxy variables". They also defined algorithmic bias as disparate impacts occurring "when a facially neutral policy or practice has a disproportionate adverse effect on a protected group", which can be extrapolated to all marginalized groups. The risks and benefits posed to society at-large warrant further exploration of these impacts of algorithmic political bias due to the rapid development of algorithms, including those underlying social media.

As social media algorithms are trained on existing social media content, this poses the potential to exacerbate discrimination against marginalized communities. Wang, et al. noted that China, Japan, and South Korea "have also recognized the importance of addressing algorithmic discrimination". China's New Generation Artificial Intelligence Development Plan called for new laws and regulation "to ensure the safe and responsible use of AI" including prioritizing the prevention of discrimination, while Japan's AI Utilization Guidelines "emphasize the need for

fairness, accountability, and transparency in AI systems." Nations with more cultural similarities to the United States, such as Canada and Australia, had also, at the time, proposed bills, such as Canada's Algorithmic Accountability Act (Bill C-27) and Australia's AI Ethics Framework, to move towards this goal (2024). Some proposed bills in the U.S. Congress, in recent years, have been receptive to the possibility of preventing discrimination and algorithmic harm, most notably: House Resolution 4624 by the U.S. House of Representatives during the 118th Congress.

H.R. 4624: The Would-be Solution for Algorithmic Transparency in the United States.

In the 2023-2024 session of the U.S. House of Representatives, H.R. 4624 was proposed "to prohibit the discriminatory use of personal information by online platforms in any algorithmic process", requiring transparency for algorithmic processes using personal data. The bill describes an algorithmic process as "a computational process… that processes personal information or other data for the purpose of determining the order or manner that a set of information is provided, recommended to, or withheld from a user of an online platform". Such actions include commercial content, social media posts, and other content involved in "automated decision making, content selection, or content amplification" processes.

The bill also proposed that oversight would be provisioned by a taskforce involving the Federal Trade Commission, several federal departments, the Consumer Financial Protection Bureau, the Federal Communications Commission, the Federal Elections Commission, and the White House Office of Science and Technology Policy. In doing so, this task force would have conducted studies on discrimination based on algorithmic bias, using personal data. This task force would have also imposed civil consequences, including "injunctive relief, declaratory

relief, damages, civil penalties, restitution, and any other relief the court deems appropriate." Specifically, relief would have included: the greater of $2,500 or the actual cost of damages, punitive damages, the costs of litigation, and injunctive or declaratory relief (118th Cong., 2023).

This was an inherently consequential proposal where oversight would have provided for meaningful review of cases, while the threat of and imposition of civil consequences would have provided redress for victims of algorithmic discrimination (given that the courts side with their claim). Crucial aspects of this policy also have a basis in research. For instance, social media algorithms use similar interests to aggregate users into echo chambers, via clustering, where these users are encouraged to stay within their clustered in-group. In many cases, these bubbles radicalize users who did not initially hold polarized beliefs (Baumann et al., p. 1, 2020). In doing so, companies have leveraged user data to develop algorithms promoting political polarization.

And, in 2018, Dafoe anticipated the potential future advances in AI technologies as they pertain to algorithms and warned of sociopolitical challenges that could be exacerbated by AI technologies. Thus, he proposed an "AI ideal governance model" architected and enforced by "governance institutions… capable of providing security, ensuring safety from non-aligned AI", and "stabilizing technological development" to prevent evolving risks. Developing and implementing ethical AI policy in this manner could encourage transparency and reduce algorithmic political bias. This proposal also integrates well with Lessig's four modalities of laws, norms, market, and architecture, cited by Tavani, which provide an extensive framework for cyberspace regulation (p. 240-241, 2016). Architecting regulatory bodies, (e.g. the task force in H.R. 4624), regulatory market and norms penalties (e.g. public campaigns to apply pressure

and regulatory fines), and laws, (e.g. GDPR and H.R. 4624) can hold bad actors accountable. These also provide opportunities for cooperation among the government, the private and non-profit sectors, and private citizens to normalize algorithmic transparency in the modalities of law, norms, market, and architecture which, in turn, can mitigate algorithmic political bias.

Critique and Unintended Potential Consequences

The Challenges of Implicit Bias.

Wang, et al. also described "unintentional discrimination in algorithmic systems "as a "complex issue that requires careful attention and robust methods for detection and regulation." They also explained that unintentional discrimination "arises from biases embedded in the data, features, or models used by the algorithm", without explicit intent. This poses significant challenges to gaining "a deep understanding of the algorithm's inputs, processing, and outputs" and relevant social and historical context. Thus, efforts to mitigate algorithmic political bias may implicitly introduce other bias. For instance, if a team working to mitigate algorithmic political bias lacks the relevant and historical context for a specific issue, they may implicitly introduce their own biases. Wang, et al. also stated how detection is only the first step towards algorithmic fairness as many systems with embedded biases already exist, highlighting the necessity of regulation and mitigation. Feedback loops also present additional challenges as methods, such as gradient boosting and neural networks, use outputs from previous layers as inputs for subsequent layers, rendering it difficult to easily mitigate issues via reverse engineering.

Wang, et al. suggested using regulatory approaches to "provide clear explanations of how individual decisions are made." Approaches include applying statistical analysis on outcomes to

help determine an algorithm's fairness and algorithmic audits, involving "a systemic examination of the algorithm's design, implementation, and use to identify potential sources of bias." Impact assessments "to proactively assess the potential discriminatory effects of their algorithmic systems" and other "identified risks" can also provide value by interfacing with members of marginalized communities for feedback on their concerns and perspectives. Requiring private and public sector entities to regularly monitor and report on algorithmic system outcomes, mandating bias detection and mitigation in development, and mandating interpretable appeals processes can enhance algorithmic transparency. Finally, the Algorithmic Justice League (AJL) were cited as "an organization that combines research, policy advocacy, and public engagement to raise awareness of algorithmic bias and develop strategies for mitigating it." As the AJL has driven public pressure and accountability, increasing public awareness as a result (2024). Integrating public input, the non-profit and private sectors, and the government can improve the effectiveness of policy to mitigate algorithmic political bias through increased public awareness.

Duty and Virtue Ethics Towards Algorithmic Fairness Interventions.

Weerts, et al. emphasized that, while "EU discrimination law seemingly draws a clear line between the obligation to avoid both direct and indirect discrimination and the mere possibility to adopt positive action measure", taking "more active steps to avoid the replication, amplification, or even the introduction of inequalities" is necessary. However, at issue is the fact that this delineation may not be so clear cut in many cases and is usually context specific, often leading to fairness interventions raising "the question of whether a negative obligation to refrain from discrimination is sufficient and appropriate in the context of algorithmic decision making." Weerts, et al. also promoted "moving towards a positive obligation to prevent discrimination in

algorithmic processing" may be preferable in provisioning risk management obligations for data governance, technical documentation, and transparency and accuracy, relying on duty-based, conscientious incentives in addition to consequential disincentives. They also suggested that "a positive obligation could introduce a duty to justify the use of a target variable that is associated with a protected characteristic" and "have the power to truly transform the baseline on which 'neutrality' is premised to break the circle of historical inequalities" (2024).

While these are valid points, the reality is that proactive measures, such as algorithmic fairness interventions or an AI engineer's explicit duty to ethically deal with data, must use external consequence as internal motive(s) may lead companies or political organizations to use data unethically and disregard said obligation to protect the public. At the same time, a focus on duty, character or virtue unfairly burdens individuals when companies and organizations should take on a collective responsibility for the embedded, implicit and explicit biases present in their data and algorithms. The appropriate consequences must also be imposed to protect private user data and the public at-large. An emphasis on individual accountability is also problematic as it frames algorithmic bias in terms of individual misdeed rather than as a societal issue.

Legal Sanctions May Not Be Enough.

Finally, Mendez-Suarez, et al. stated how "the application of legal sanctions as punishment for organizational misconduct is not deterring wrongdoing" as ineffective AI regulation and a lack of public accountability remain issues (Mendez-Suarez et al., p. 215-216, 2023) if AI remains uninterpretable. This points to unintended consequences, such as the fact that inconsequential fines may result in public assumptions that entities that are wealthy enough

can simply “pay to play”, rather than facing meaningful consequences. This is another reason for why awareness, such as those advocated for by the AJL, are vital in increasing public awareness to shame public perception and increase public awareness and collective action. Regardless, this does not negate the fact that such laws increase public awareness of this issue. Simultaneously, AI policy is an underdeveloped field of study, so opportunities are still abundant for AI policy to evolve. Such opportunities include Gervais’s suggestion to write “norms in a technologically neutral fashion and interpret them dynamically” with an ‘“observatory” function” reporting on implementation. This would result in interpretable, global standards for AI policy to “shed light on implementation” by providing frameworks for other jurisdictions to follow (p. 404, 2021).

Conclusion: Consequentialist vs Duty, Character and Virtue Actions

Despite the hurdles against algorithmic transparency and fairness policy, consequential actions are necessary. Even if ineffective, they promote legal and societal norms that emphasize using policy to mitigate algorithmic harm, implicitly leading citizens to keep algorithmic political bias at top of mind. While they have their place, morality based on duty and virtue explicitly requires individual buy-in, whereas consequence is best applied through collective means – in this case, government or another regulatory body. It cannot be understated that personal and private and public sector self-governance based on duty, character, and virtue are vital to enforce norms, market forces, and the regulatory architecture to maintain algorithmic transparency and fairness. Still, regulatory actions to enforce algorithmic transparency require consequential action to ensure that those who engage in unethical behavior are held accountable.

References

Algorithmic Justice and Online Platform Transparency Act, H.R. 4624, 118th Cong. (2023). https://www.congress.gov/bill/118th-congress/house-bill/4624/text

Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., & Starnini, M. (2020). Modeling Echo Chambers and polarization dynamics in Social Networks. *Physical Review Letters, 124*(4). https://doi.org/10.1103/physrevlett.124.048301

Bezzubova, E. (2018, January 26). Digital Depersonalization. Psychology Today. https://www.psychologytoday.com/us/blog/the-search-self/201801/digital-depersonalization

Dafoe, A. (2018). AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 1442*, 1443.

Gervais, D. J. (2021). Towards an effective transnational regulation of AI. *AI & SOCIETY, 38*(1), 391–410. https://doi.org/10.1007/s00146-021-01310-0

Kelly, M. (2021, December 9). *Political polarization and its Echo Chambers: Surprising new, cross-disciplinary perspectives from Princeton*. Princeton University. https://princeton.edu/news/2021/12/09/political-polarization-and-its-echo-chambers-surprising-new-cross-disciplinary

Klee, M. (2025, May 16). *Grok pivots from "white genocide" to being "skeptical" about the Holocaust*. Rolling Stone. https://www.rollingstone.com/culture/culture-news/elon-musk-x-grok-white-genocide-holocaust-1235341267/

Leingang, R., & Gedeon, J. (2025, June 17). How the right spread "brutal and cruel" misinformation after Minnesota lawmaker killings. The Guardian.

https://www.theguardian.com/us-news/2025/jun/17/minnesota-lawmaker-killings-misinformation-rightwing

Méndez-Suárez, M. (2023). Do current regulations prevent unethical AI practices? *Journal of Competitiveness, 15*(3). https://doi.org/10.7441/joc.2023.03.11

Peters, U. (2022). Algorithmic Political Bias in Artificial Intelligence Systems. *Philosophy & Technology*. *35*(25). https://doi.org/10.1007/s13347-022-00512-8

Shapiro, E., & Margolin, J. (2025, June 15). Minnesota assassination suspect Vance Boelter allegedly had dozens of Democrats on a list. ABC News. https://abcnews.go.com/US/gov-walz-rep-omar-dozens-minnesota-democrats-gunmans/story?id=122847427

Tavani, H. T. (2016). *Ethics and technology* (5th ed.). John Wiley & Sons, Incorporated.

Veluru, C. S. (2024). Responsible artificial intelligence on large scale data to prevent misuse, unethical challenges and security breaches. *Journal of Artificial Intelligence & Cloud Computing*, 1–6. https://doi.org/10.47363/jaicc/2024(3)331

Wang, X., Wu, Y. C., Ji, X., & Fu, H. (2024). Algorithmic discrimination: Examining its types and regulatory measures with emphasis on us legal practices. Frontiers in Artificial Intelligence, 7. https://doi.org/10.3389/frai.2024.1320277

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. (2024). The neutrality fallacy: When algorithmic fairness interventions are (not) positive action. The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2060–2070. https://doi.org/10.1145/3630106.3659025

Appendix

No generative AI was sourced for the contents of this document in its current iteration.