Examining Algorithmic Bias in Recommender Systems: Ethical Challenges and Solutions

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

Eli S. Herman

Spring 2025

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

Sean Murray, Associate Professor of STS, Department of Engineering and Society

Introduction

AI-driven recommender systems have become integral to how people discover information and products in various domains, from entertainment to e-commerce. These systems offer personalized suggestions intended to enhance user satisfaction and engagement. However, a growing body of evidence shows that recommender algorithms often suffer from algorithmic bias, which can limit consumer choice and reinforce societal inequities. For example, Safiya Umoja Noble's research revealed that a simple Google search for terms like "Black girls" returned predominantly pornographic sites, illustrating how ostensibly neutral algorithms can perpetuate racist and sexist biases. Noble concluded that Google's search algorithm had "failed the public, particularly people who are oppressed [and] under-represented" in its presentation of results. Similarly, Latanya Sweeney (2013) found that online advertising algorithms displayed arrest record ads far more often for Black-identifying names compared to white-identifying names, reflecting a discriminatory pattern in automated decision-making. Such examples underscore that algorithmic systems are not inherently objective—"algorithms are opinions embedded in code," as mathematician Cathy O'Neil famously observed. The biases of designers, or the skewed data these systems learn from, can result in outputs that disadvantage certain groups.

In the context of recommender systems, algorithmic bias often manifests through feedback loops and popularity bias. Algorithms that prioritize popular or mainstream items tend to keep recommending what most users already see or hear, thereby marginalizing niche or alternative options. This creates a filter bubble or "echo chamber" effect: users are exposed to an increasingly narrow range of content reinforcing their existing preferences. Over time, the system's focus on prevalent patterns can amplify existing social biases. For instance, major

content platforms like Netflix and Spotify have faced criticism for recommendation engines that favor well-known content and overlook diverse or minority voices. These personalized services can inadvertently limit exposure to new genres or creators. As one article on Spotify's algorithmic design explains, "the algorithm's very prowess in crafting personalized recommendations could inadvertently perpetuate existing biases, disproportionately spotlighting certain genres, artists, or cultural trends" (Torabi, 2023). Likewise, Netflix's algorithms optimize for engagement by promoting popular or self-produced titles, which can make it difficult for niche, foreign, or independent films to surface. As Martin Scorsese put it, "Algorithms, by definition, are based on calculations that treat the viewer as a consumer and nothing else"—a paradox attributed to algorithmic curation that highlights mainstream content while burying the rest. As Spandana Singh explains, "Platforms such as Amazon and Netflix produce films and television shows based on behavioral data collected on their users through these systems… shaping the database of options that users have to choose from" (Singh, 2020, p. 6). The net effect is that users' choices are subtly steered by systems designed for stickiness and scale, rather than diversity or cultural exploration. Figure 1 offers a helpful lens for understanding how bias can become embedded in algorithmic systems, from data collection to deployment, reinforcing the cultural narrowing described above.

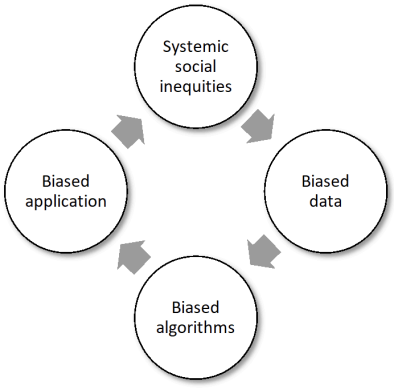

Figure 1. *A framework for identifying and mitigating bias in algorithmic systems, originally developed for mental health apps. Adapted from Timmons (2023).*

Against this backdrop, this research examines the ethical challenges posed by algorithmic bias in recommender systems and explores solutions to make these systems fairer and more accountable. This inquiry is pursued alongside a technical capstone project called Pairings, which involves building a machine-learning-based wine recommendation app. The Pairings app is designed as a case study in mitigating bias: it personalizes wine suggestions based on meal context, taste preferences, and budget, while explicitly aiming to avoid the narrow feedback loops seen in other recommenders. By combining content-based filtering (matching wine characteristics to meal attributes) with collaborative filtering (learning from user ratings), Pairings implements a hybrid model intended to balance personalization with diversity. Prior research supports this approach as Fernández-Tobías et al. (2011) demonstrated that hybrid recommender models can improve accuracy by leveraging multiple algorithms, reducing the risk of one-sided recommendations. In aligning the technical design with ethical considerations, the project seeks to ensure that the system's accuracy and user satisfaction goals do not come at the cost of reinforcing bias. This dual focus sets the stage for a broader discussion on how recommender systems can be designed and evaluated not just for efficiency, but also for fairness, transparency, and accountability.

Algorithmic Bias and Its Impacts on User Choice

Algorithmic bias in recommendation systems raises significant ethical and societal concerns by influencing user behavior and potentially entrenching existing inequalities. When

recommendation algorithms prioritize what is already popular or cater to majority tastes, they can create a self-perpetuating cycle: popular items get more exposure, become more popular, and crowd out less common alternatives. Over time, this cycle (an algorithm-driven feedback loop) can narrow a user's content diet. Users receive recommendations that reinforce familiar preferences or widely held trends, while content outside the mainstream—be it music by emerging artists, niche genres of films, or products from minority-owned businesses—remains hidden. This not only reduces individual choice and serendipitous discovery, but also has broader cultural implications: the perspectives and contributions of minority groups may be underrepresented in what people watch, read, or buy.

Research by Noble (2018) and others illustrates that biased algorithms can replicate societal power imbalances under the guise of personalization. Noble's critique of search engines showed how algorithms "embed societal biases into their rankings," disproportionately associating women of color with hypersexualized or negative content. In a parallel vein, O'Neil (2016) documents how data-driven algorithms used in domains like credit scoring, hiring, or criminal justice can unfairly penalize marginalized communities. These Weapons of Math Destruction, as O'Neil calls them, tend to be opaque, unregulated, and scalable—factors that allow biased outcomes to proliferate widely without easy detection. Although recommender systems for movies or music might seem benign by comparison, they too can have meaningful impacts on whose voices get amplified. If Netflix's algorithm systematically favors English-language or U.S.-produced content, for example, it may sideline content from other cultures, subtly shaping viewers' worldviews. According to Singh (2020), Netflix's recommendation system drives approximately 80% of the hours of content streamed on the platform, underscoring its role in shaping what users watch. However, the platform has offered

limited transparency into how its recommendation algorithms operate, raising concerns about potential bias and discrimination. Singh notes that Netflix's system influences not just what content is surfaced, but even what gets produced, as the company creates original shows based on behavioral data collected through the system. This feedback loop may narrow the diversity of options available and subtly shape viewers' media consumption in favor of mainstream or popular trends (Singh, 2020, pp. 6, 32).

The impacts of these biases are not merely theoretical. Case studies highlight concrete consequences: A Netflix user who mostly watches popular sitcoms might never be exposed to critically acclaimed documentaries that challenge social biases, because the algorithm doesn't deem them "relevant." A talented indie musician on Spotify might struggle to find an audience because the recommendation engine keeps pushing superstar artists with established listenership. Such skewed exposure can reinforce a rich-get-richer dynamic in cultural markets, where already dominant artists and franchises gain more prominence, while others are invisibly suppressed. Moreover, biased recommendations can affect user perceptions and beliefs. If someone's news feed or video suggestions consistently lean toward a single ideological perspective (due to algorithms learning and reinforcing that preference), it can create a false sense of consensus and reinforce confirmation bias. In sum, algorithmic bias in recommender systems has a twofold impact: it limits the diversity of content that individuals engage with, and it perpetuates broader patterns of inequality by giving disproportionate advantages to already advantaged content producers or viewpoints.

These issues have started to draw public scrutiny and demand for change. Users and ethicists are questioning whether personalization is coming at too high a cost to the public interest. There is a growing recognition that, left unchecked, recommender systems could

contribute to societal problems such as diminished cultural diversity, discrimination, or polarization. The next sections explore how scholars and practitioners are responding to these ethical challenges—by developing frameworks to understand fairness in algorithms, and by proposing strategies to make recommender systems more inclusive and accountable.

Ethical Frameworks for Fair Recommender Systems

To address algorithmic bias, researchers have been applying ethical frameworks and theories of fairness to the design of AI systems. A key realization is that fairness is not a one-dimensional concept; it has multiple definitions and trade-offs that need to be balanced in algorithmic contexts (Binns, 2018a). Reuben Binns (2018a) draws on political philosophy to examine different notions of fairness in machine learning, highlighting that what one considers "fair" can depend on moral values and context. For instance, an algorithm could be fair in the sense of treating everyone the same (formal equality), yet still produce unequal outcomes for historically disadvantaged groups. In recommender systems, this might translate to giving every user the "same chance" to have their content recommended—yet if the underlying data reflects existing inequalities (e.g. fewer ratings for minority-created content), the outcomes can still be biased. An important lesson is that designers must clarify which fairness criteria they aim to satisfy and acknowledge the limitations of any single metric.

Beyond abstract definitions, scholars emphasize practical principles like transparency, accountability, and inclusion in guiding the development of fairer systems. Binns et al. (2018b), in a study of people's perceptions of algorithmic decisions, found that users value transparency and justification when algorithms affect them. Lack of insight into how a recommendation was made can lead to feelings of indignity or helplessness. In interviews, one participant described an

opaque algorithmic decision by saying “it’s just simply reducing a human being to a percentage. It’s not taking any of [their] actual ability… into account.”. This sentiment captures the ethical imperative for user-centered design: algorithms should be understandable and should treat users as more than just data points. If a recommendation system can explain, even in simple terms, why it suggested a certain item (“because you listened to X, you might like Y”), users are more likely to perceive it as fair or at least acceptable, even when they choose not to follow the advice. Transparency is also linked to accountability: if stakeholders can inspect how an algorithm works or on what data it is trained, they can better identify biases and demand corrections. Nick Diakopoulos (2016) argues that as algorithms play a larger role in society, we must demand they be accountable to the public—meaning there should be mechanisms to audit and contest their decisions, much as we do with human decision-makers. In the context of recommender systems, accountability might involve regularly evaluating recommendation outcomes for bias, allowing users to flag problematic suggestions, or enabling external reviews of the recommendation algorithms for fairness issues.

However, researchers like Barocas, Hardt, and Narayanan (2019) caution that transparency alone is not a panacea. Simply revealing how an algorithm works does not automatically fix biased outcomes—especially if the root cause is biased data. As artist Peter Gabriel once observed, while “the present is defined by freedom of choice, the future will be defined by freedom from choice”—a chilling notion when applied to algorithmic personalization. If a music recommendation algorithm is trained on listening data that underrepresents jazz or classical music, being transparent about its inner workings won’t change the fact that jazz or classical tracks might rarely be recommended. Thus, an ethical framework for recommender systems also calls for diversity in data and training. Barocas et al. (2019) advocate for actively

curating datasets to include a wide range of user demographics, content types, and cultural contexts. In practical terms, this could mean ensuring that a movie recommender's training data isn't drawn solely from one country or age group, or that a job recommendation platform's data isn't skewed by past discriminatory hiring practices. By broadening the data and continually monitoring for unequal representation, developers can make algorithms less prone to inheriting yesterday's prejudices.

Another facet of fairness is user agency. Some propose giving users more control over their recommendation settings—allowing them to dial up novelty or diversity, for example, or to opt out of certain personalization aspects. This stems from the idea that fairness can be enhanced by respecting individual users' goals and giving them a say in how the algorithm serves them (Binns et al., 2018b). If a user feels a recommendation list is too narrow, they might toggle a feature to get more varied suggestions, thereby breaking out of the filter bubble. While not all users will utilize these options, their availability is a sign that the system designers recognize a responsibility to accommodate different definitions of a "good" recommendation.

In summary, several ethical guidelines emerge for recommender systems: make the algorithms transparent and their outcomes justifiable to those affected; ensure accountability through oversight and avenues for redress; use diverse, representative data to train models; and incorporate user-centered design that respects individuals' values and autonomy. These principles set the stage for concrete strategies to operationalize fairness in recommendation engines.

Strategies for Mitigating Bias in Recommender Systems

Building on the above frameworks, researchers and practitioners have proposed a variety of strategies to identify, mitigate, and prevent bias in recommender systems. These strategies span the pipeline of system development—from data preprocessing to algorithm design to interface tweaks—and often need to be used in combination to be effective. A recent comprehensive survey by Mehrabi et al. (2021) notes that bias in machine learning can be tackled at multiple levels, such as by preprocessing data to remove or balance biased representation, by introducing in-processing constraints or regularizers to make models treat groups more equally, and by post-processing outputs to ensure fair distribution of results. In the context of recommendations, one preprocessing step could be augmenting the training dataset with additional examples of under-recommended items so that the algorithm learns a richer mapping of user preferences. This approach addresses the problem at its root—if the data is less biased, the learned model is less likely to be biased.

During the algorithm training phase, another strategy is to adjust the objective function to value not just accuracy or click-through rates, but also diversity and fairness metrics. For example, a recommender algorithm could be penalized if its top-10 suggestions for a user are too homogenous (for example, all are blockbuster action movies), thereby encouraging it to include a couple of less obvious, diverse picks. Some research prototypes implement fairness-aware recommendation algorithms that explicitly re-rank or filter results to improve exposure for items or creators from underrepresented categories. One challenge here is to balance fairness with personalization: the goal is not to show every user a perfectly demographically representative set of content, but rather to widen the scope of recommendations just enough to avoid unjust exclusion of certain content. Techniques like relevance calibration can help, where the

recommender produces a baseline list of personalized items and then injects additional items that increase diversity while still being reasonably relevant to the user's interests.

Another mitigation approach is algorithmic transparency and user feedback loops. By explaining recommendations and exposing potential biases, systems can invite corrective feedback. For instance, if a news app tells a user "You are seeing more political articles because you frequently read politics news," the user might realize they are in a bubble and decide to explore other sections—or conversely, they might tell the app they actually want less of that category. In both cases, giving insight can prompt adjustments that reduce bias. Some platforms have started to implement tools for this; YouTube and Facebook, for example, allow users to "tune" their feed by selecting content they want to see more or less of. Though far from perfect, these features acknowledge that the one-size-fits-all algorithmic ranking can fail for some users or impose values that the user might not share. User feedback thus becomes part of the solution: algorithms can be retrained or updated based on aggregate signals that users provide about recommendation quality. To visualize this cycle, researchers have illustrated, in Figure 2, how user interactions feed back into the recommendation model itself, creating a loop that can either reinforce or disrupt existing biases (Chaney, Stewart, & Engelhardt, 2018).

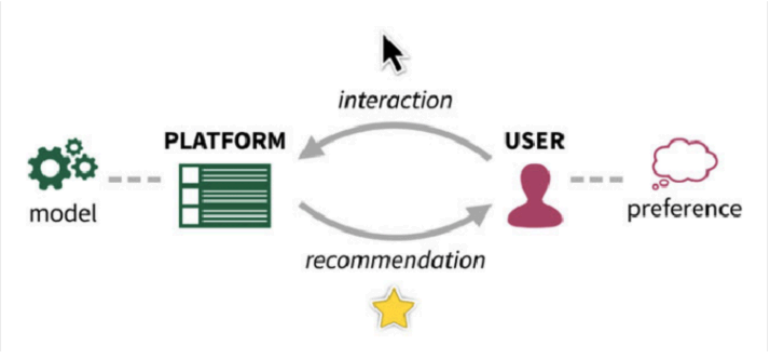

Figure 2. *The recommender system feedback loop. As users interact with content, their behavior updates preferences, which feeds back into the model, shaping future recommendations. Reproduced with permission from Chaney, Stewart, and Engelhardt (2018).*

Importantly, the organizational context and policy also play a role in mitigating bias. Tech companies can institute regular fairness audits of their recommender systems—evaluating, for example, whether certain artists or genres are consistently under-recommended relative to their actual quality or popularity in niche communities. There have been calls for industry standards or regulatory guidelines to ensure platforms disclose how their recommendation algorithms work and what measures they take to prevent discrimination or unfair treatment of content providers. Just as consumer protection laws guard against deceptive advertising, one could envision algorithmic accountability regulations that require companies to assess and report biases in automated recommenders (Diakopoulos, 2016). While such regulations are still nascent, some jurisdictions are discussing transparency requirements for high-impact algorithms, which could include those that curate media or job opportunities.

In academic research, new methods are emerging to measure recommender system bias. One concept is calibrated recommendations, which aims to ensure that the mix of suggestions a user receives aligns with the diversity of their own past preferences. For instance, if a user historically listens to 70% rock and 30% jazz, a calibrated recommendation list would try to include roughly 30% jazz suggestions. This respects the user's demonstrated interests in a balanced way, preventing the system from completely overwhelming a minority interest with the majority one. If a user's tastes themselves are biased or narrow, calibrated output won't broaden

them beyond what the user has signaled; however, it avoids the problem of the algorithm exaggerating biases. Another method is counterfactual evaluation – asking how the recommendations would differ if some aspect were varied. This can reveal if an algorithm is relying too heavily on a factor that could be bias-laden. If, in a counterfactual scenario, an equally good recommendation set can be produced that includes more diverse items, it suggests the original algorithm might be unnecessarily skewed.

The Pairings wine recommendation app exemplifies a couple of these strategies in practice. In developing Pairings, I prioritized dataset diversity by including a broad range of wines so that the model doesn't just learn to recommend the most commonly rated wines. The hybrid recommendation approach it uses is also a bias-mitigation tactic: content-based filtering ensures that even less-known wines can be suggested if they objectively match the user's meal context, while collaborative filtering adds the wisdom of crowd preferences. During testing, the developers are monitoring the recommendations to see if the algorithm disproportionately favors certain wine varieties or neglects others; the model is tweaked accordingly to improve balance. This iterative approach reflects a general principle for all recommender systems: bias mitigation is not a one-off fix but an ongoing process. As user behavior and content catalogues evolve, so too must the strategies to keep recommendations fair and inclusive.

Conclusion

Algorithmic bias in recommender systems is a complex problem at the intersection of technology and society. As this research has discussed, the personalization offered by modern recommenders comes with the risk of reinforcing existing preferences and societal biases, which can limit the diversity of content that users see and amplify inequities in visibility and

opportunity. The ethical challenges span issues of fairness, transparency, and accountability. Addressing these challenges requires a multi-faceted effort. On one hand, technical solutions—such as more diverse training data, fairness-aware algorithms, and user controls—can reduce bias in recommendations. On the other hand, ethical design principles and possibly regulatory oversight are needed to guide the development and deployment of these algorithms in alignment with societal values.

The exploration of Netflix and Spotify's recommendation practices, alongside scholarly critiques, shows that without conscious intervention, recommender systems will naturally favor the status quo or the majority, often to the detriment of minority interests. Yet, the case studies and strategies outlined also provide reason for optimism: we have the knowledge and tools to design recommender systems that balance personalization with fairness. By diversifying the content that algorithms consider and exposing users to a broader array of options, we can enrich user experience rather than constrict it. By implementing transparency and accountability, we can build greater trust in algorithmic systems and ensure they serve the public good, not just corporate or majority interests.

The Pairings app developed in the accompanying technical project offers a tangible example of how these ideas can be put into practice. It demonstrates that even a niche application—a wine recommender—benefits from integrating fairness considerations from the ground up. The app's hybrid algorithm and design choices were made not only for accuracy, but also to prevent the kind of bias-driven limitations seen in other platforms. Early results from Pairings suggest that it is possible to provide personalized recommendations that delight users while still introducing them to a variety of choices they might not have discovered otherwise. In

doing so, it validates the notion that personalization and diversity need not be mutually exclusive goals.

In conclusion, tackling algorithmic bias in recommender systems is both an ethical imperative and a path to better technology. Recommender systems wield significant influence over what information and entertainment people consume; ensuring those systems are fair and inclusive will help create a more equitable digital ecosystem. Continued collaboration between technologists, ethicists, and policymakers will be crucial to refine these solutions and address new biases that emerge. By striving for fairness and accountability, we not only improve the algorithms themselves but also uphold the values of diversity and justice in the increasingly AI-driven society.

References

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved from fairmlbook.org.

Binns, R. (2018a). Fairness in machine learning: Lessons from political philosophy. In Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (pp. 149–159). New York, NY: ACM.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018b). "It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Paper 377, 1–14). New York, NY: ACM.

Brode, M. (2020). Machine learning for wine recommendations: A case study. Journal of Food and Beverage Analytics, 7(3), 145–158.

Bucher, S. (2024, January 21). Algorithmic symphonies: How spotify strikes the right chord - USC viterbi school of engineering. USC Viterbi School of Engineering - USC Viterbi School of Engineering. https://illumin.usc.edu/algorithmic-symphonies-how-spotify-strikes-the-right-chord/#:~:text=transparency%20of%20how%20this%20data,trending%20or%20popular%20artists%2C%20causing

Chaney, A. J. B., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 224–232). https://doi.org/10.1145/3240323.3240370

Diakopoulos, N. (2016). Accountability in algorithmic decision-making. Communications of the ACM, 59(2), 56–62.

Martinez, S. (2022, August 16). Streaming service algorithms are biased, directly affecting content development. AMT Lab @ CMU. https://amt-lab.org/blog/2021/11/streaming-service-algorithms-are-biased-and-directly-affect-content-development#:~:text=the%20popularity%20to%20enter%20the,"

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York, NY: NYU Press.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York, NY: Crown.

Ortiz, S. (2024, July 19). LWL #25 discrimination in data and artificial intelligence - data-pop alliance. Data.

https://datapopalliance.org/lwl-25-discrimination-in-data-and-artificial-intelligence/#:~:text=LWL%20,In

Resnick, P., & Varian, H. R. (1997). Recommender systems. Communications of the ACM, 40(3), 56–58.

Rollenhagen, L. (2021, May 10). For Sweeney, data is never bias-free. Osano. https://www.osano.com/articles/meet-latanya-sweeney#:~:text=likely%20to%20be%20related%20to,name%20would%20be%20 searched%20online

Rollmann, R. (2018, January 30). Don’t google it! how search engines reinforce racism " popmatters. PopMatters. https://www.popmatters.com/algorithms-oppression-safiya-umoja-noble-2529677349.html#:~:text=That’s%20patently%20wrong%2C%20argues%20Noble,It%20doesn’t%20provide%20you%20real

Singh, S. (2020, March 25). Why am I seeing this? The case for greater transparency around the algorithms that curate our digital experiences. New America. https://www.newamerica.org/oti/reports/why-am-i-seeing-this/

Smith, J., & Allen, D. (2019). Recommender systems in gastronomy: Bridging user experience and AI. International Journal of AI in Hospitality and Tourism, 6(1), 40–55.

Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the ACM, 56(5), 44–54.

Timmons, A. (2023). Bias feedback loop in mental health algorithms. In AI bias in mental health apps. UT Austin College of Liberal Arts. https://liberalarts.utexas.edu/psychology/news/research-paper-from-adela-timmons-looks-at-ai-bias-in-mental-health-apps

Torabi, N. (2023, August 28). The inner workings of Spotify’s AI-powered music recommendations: How Spotify shapes your playlist. Medium. https://medium.com/beyond-the-build/spotify-recommendation-system-nima-torabi-aug2023

Verbert, K., Manouselis, N., Ochoa, X., Wolpers, M., Drachsler, H., Bosnic, I., & Duval, E. (2012). Context-aware recommender systems: A review and future directions. User Modeling and User-Adapted Interaction, 22(1–2), 157–190.