Proceedings of the 4th International Conference on Global Politics and Socio-Humanities
DOI: 10.54254/2753-7048/2026.34114

Gender Stereotypes under Social Media Algorithms: Mechanisms, Legal Challenges, and Regulatory Pathways

Yutong Zhang

Department of Sociology, University of California, Santa Barbara, United States
yutong904@ucsb.edu

Abstract. Social media platforms have increasingly delegated content curation to algorithmic recommendation systems, and the gender consequences of this delegation deserve more legal attention than they currently receive. Rather than simply reflecting pre-existing stereotypes, these systems actively reinforce them: through echo chambers that progressively narrow the gender representations users encounter; through automated gender classification that sorts users into binary categories before they have done anything at all; and through training data that carries historical bias forward into new models. Each of these mechanisms is empirically documented. None of them is adequately addressed by existing law. Platforms claim neutrality; privacy protections make bias investigation legally difficult; and the diffuse, cumulative nature of stereotyping harm does not fit what traditional anti-discrimination categories were designed to handle. This paper draws on comparative regulatory analysis across the EU, the United States, and China, and on recent legislative developments including the AI Act, to propose three targeted interventions: mandatory algorithmic auditing for gender bias, legal requirements for gender-inclusive design in recommendation architecture, and transparency obligations that give users meaningful information about why they see what they see. The goal throughout is to identify regulatory responses that are proportionate and feasible, not merely aspirational.

Keywords: algorithmic recommendation systems, gender stereotypes, algorithmic bias, social media regulation, EU AI Act

1. Introduction

Social media platforms were once commonly described as mirrors of society, spaces where existing behaviors, interests, and inequalities were reflected back to users without significant distortion. This framing was always somewhat flattering to the platforms. But by now it has become genuinely inaccurate. The systems that determine what content users see on TikTok, Instagram, Facebook, or Weibo do not passively reflect anything. They actively select, rank, and deliver material based on continuous behavioral inference, and the choices they make have measurable effects on how users understand gender roles, behavioral norms, and what kinds of lives are available to people like them.

The evidence is not subtle. A 2023 experiment conducted by Global Witness found that Facebook job advertisements for mechanic positions were delivered to audiences that were 91% male, while advertisements for pre-school teaching vacancies reached audiences that were 79% female [1]. The

© 2026 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: 10.54254/2753-7048/2026.34114

advertisers had not specified any demographic targeting. The platform made those distributions happen on its own. The European Parliament's 2023 study on social media and women found consistent patterns across TikTok and Instagram of career and finance content being directed toward male-classified users while domestic, beauty, and care content was directed toward female-classified users [2]. Neither finding is particularly surprising in isolation. But both point to systems that are not merely amplifying user preferences; they are constructing and reinforcing gendered content environments in ways that users cannot see and have little power to challenge.

An algorithmic recommendation system, as used here, is the automated process through which a platform selects, ranks, and delivers content to individual users based on inferred preferences, behavioral data, and demographic characteristics. Gender stereotypes in the digital context refers to the systematic assignment of attributes, roles, and behavioral expectations to individuals based on their real or inferred gender, in ways that flatten individual complexity and reinforce binary, hierarchical conceptions of what men and women are supposed to want and do [3].

The connection between these two phenomena occurs at three levels. At the content level, algorithms prioritize material that has historically attracted engagement from users of a given inferred gender, producing feedback loops that intensify gender-typical content over time. At the classification level, platforms assign gender labels to users using inference systems built on binary models that cannot accommodate the diversity of gender identity. At the data level, training datasets drawn from historical behavior carry forward the biases of previous periods. None of these mechanisms requires deliberate discriminatory intent. All three produce discriminatory effects.

This paper works through three questions: what mechanisms produce the observed gender effects; why existing law cannot adequately address them; and what regulatory changes could help without disproportionate costs to expression or privacy. The analysis draws on empirical research about how platforms actually function, comparative legal analysis across the EU, the United States, and China, and recent legislative developments including the GDPR and the EU AI Act. The aim is not to identify a jurisdiction that has already solved the problem but to locate the most useful tools and think through how they might be combined and extended.

2. Three mechanisms

2.1. Echo chambers and the gradual narrowing of what users see

Jamieson and Capella introduced the echo chamber concept to describe information environments where users consistently encounter content confirming their existing beliefs while being insulated from alternatives [4]. Similar logic underlies Pariser's notion of the filter bubble, in which personalization algorithms cocoon users within a self-reinforcing information environment [5]. The mechanism translates directly to gender. When recommendation algorithms optimize for engagement, they tend to serve content that is familiar and predictable. For users with gender-typical engagement patterns, this means progressively more gender-typical content, which reinforces those patterns, which generates more gender-typical recommendations. The loop does not require any deliberate stereotyping decision at any point.

This is not speculative. Cinelli et al. documented homophilic clustering effects across Facebook, YouTube, Twitter, and Gab, showing that platform architecture consistently shapes the information diet of different user groups [6]. A systematic review covering thirty studies across Facebook, YouTube, TikTok, Instagram, and Weibo reached similar conclusions about the structural tendency of algorithms to reduce viewpoint diversity [7]. The pattern shows up early in a user's platform experience. When a new user joins, the recommendation system has little individual data and often

DOI: 10.54254/2753-7048/2026.34114

relies on demographic inference, including inferred gender, to seed initial suggestions. If a user classified as female first engages with content about cooking or personal care, the algorithm reads this as a preference signal. The feed narrows from there. The algorithm is not reflecting a preference; it is constructing one.

2.2. Automated gender classification

There is a layer beneath content recommendation that receives less attention: the automated classification of users into gender categories using behavioral inference. These Gender Classification Systems (GCSs) analyze posting patterns, content preferences, language use, and network connections to assign a gender label. That label then shapes advertising delivery, content curation, and other platform functions. Most users are unaware this classification is happening.

The technical problems are documented. Research based on a survey of over 1,600 users on X found that women were misclassified more often than men, and that LGBTIQ+ individuals experienced significantly higher misidentification rates than cisgender users [8]. This is not a random error distribution. GCSs are trained on historical data that encodes a binary model of gender, and they are structurally less accurate for users whose identities or behaviors fall outside that binary. Fosch-Villaronga et al. make the broader point that GCS use constitutes automated profiling with significant gender and privacy implications that existing frameworks are not built to handle [9]. The conceptual problem compounds the technical one. Classifying users into two gender categories is itself an act of stereotyping. It builds a simplifying assumption into platform infrastructure. Every downstream decision gets filtered through a classification the user typically cannot see and has no clear mechanism to contest.

2.3. Historical data and the feedback loop

Recommendation systems learn from historical data. When that data reflects gender-biased conditions, the models trained on it reproduce those conditions in their outputs. This is not a design oversight that can be patched easily; it is a structural feature of how machine learning works when trained on social data.

Research on language models has demonstrated that word embeddings trained on large corpora consistently encode gender stereotypical associations, reflecting the biases present in the underlying text [10]. Rathee et al. further demonstrated that algorithms trained on large language corpora learn to associate women with more negative consumer psychographic attributes than men because those associations are present in the training language itself [11]. A separately documented case involves Amazon's internal recruitment algorithm, which was trained on ten years of hiring data. Because the historical applicant pool was predominantly male, the system learned to downgrade resumes containing the word "women's." Amazon eventually shut the project down [12]. Social media is not hiring, but the mechanism is the same. Historical engagement data showing more male users engaging with technology content and more female users engaging with lifestyle content trains recommendation models to replicate those distributions. The biases do not originate with the algorithm, but the algorithm ensures they persist and compound.

DOI: 10.54254/2753-7048/2026.34114

3. Why existing law cannot handle this

3.1. The neutrality defense

The standard platform response to questions about gender bias in recommendations is a claim of algorithmic neutrality. The algorithm does not treat users differently based on gender, the argument goes; it responds to behavioral signals. Users choose what to engage with, and the system reflects those choices. Any observed gender asymmetries in content delivery are therefore a product of user behavior, not platform design.

This is wrong in two specific ways. First, platforms classify users by inferred gender at account creation and use that classification to seed initial recommendations before any individual behavioral data has been collected. Second, treating historically gender-biased behavioral patterns as the neutral baseline for future recommendations is not neutrality. As Barocas and Selbst observe in the context of big data, when systems are trained on historical data produced under conditions of inequality, they systematically reproduce and amplify those inequalities regardless of intent [13]. The neutrality claim works more as a legal strategy than a factual description. By characterizing their systems as passive reflectors of user preference, platforms aim to avoid having their conduct classified as discriminatory and to resist any affirmative obligation to pre-screen for bias.

3.2. Privacy law and anti-discrimination work against each other

Detecting gender bias in recommendation systems requires access to data about how those systems classify users and how recommendation patterns vary across the user population by gender. This is precisely the data that privacy law protects most stringently. The conflict is not incidental; it is structural, and it has no easy resolution.

Gender is not listed as a special category under Article 9 GDPR, which provides heightened protection for health data, racial origin, sexual orientation, and similar characteristics. But inferring gender from behavioral patterns may in practice reveal information about sexual orientation or health, which would bring it near the Article 9 threshold. Wachter and Mittelstadt have argued that the right to reasonable inferences represents a missing dimension of data protection law that is urgently needed in the age of algorithmic profiling [14]. The EU AI Act's Article 10(5), in force since August 2024, permits processing of sensitive data specifically for bias monitoring and correction in high-risk AI systems [15]. This is genuinely useful, but how it interacts with Article 9 GDPR remains unresolved in ways that will likely need judicial clarification before regulators can rely on it confidently.

3.3. The harm does not fit existing legal categories

Even if the privacy obstacle were removed, anti-discrimination law would still struggle. The difficulty is not procedural but conceptual.

Traditional anti-discrimination law was built for discrete, identifiable acts. A refusal to hire. A denial of housing. An exclusion from service. It presupposes a claimant who suffered a concrete and traceable harm caused by a specific decision made by an identifiable actor. As O'Neil has documented, algorithmic systems routinely produce discriminatory outcomes at scale while evading conventional legal accountability precisely because the harm they cause is diffuse, statistical, and cumulative rather than individual and traceable [16]. Algorithmic gender stereotyping does not produce harm in any conventional legal sense; it produces a gradual narrowing of cognitive

DOI: 10.54254/2753-7048/2026.34114

exposure and a structural limitation on the range of identity possibilities users encounter over time. In the United States, Section 230 of the Communications Decency Act has historically shielded platforms from liability for recommendation outputs. The EU's Digital Services Act (2022) requires platforms to assess systemic recommendation risks and offer non-personalized alternatives, but it does not create individual remedies [17]. No jurisdiction examined here has bridged this gap.

4. What is currently being tried

4.1. Platform self-regulation

Without binding external requirements, platforms have developed internal responses to bias concerns. Community standards prohibit certain forms of gender-based harassment. Content moderation systems remove overtly discriminatory material. Some platforms have made documented adjustments to recommendation outputs claimed to increase content diversity. TikTok has published materials describing filter bubble prevention efforts. Meta has described advertising delivery audit processes.

None of this is worthless, but the structural problem is a conflict of interest that internal policy cannot resolve. Engagement drives revenue. Gender-typical content generates reliable engagement because it is familiar and predictable. The financial incentive for platforms runs in the opposite direction from the goal of diversifying gender representations in recommendation feeds. Without external accountability, self-governance on this issue depends on individual teams caring enough to push against the commercial logic of their own organizations. That is an unstable foundation for consistent protection.

4.2. Comparative legal approaches

The EU has gone furthest toward binding requirements. The AI Act (Regulation 2024/1689), in force from August 2024, establishes mandatory bias auditing, training data documentation, and fundamental rights impact assessments for high-risk AI systems [18]. Cerezo-Martínez et al. have analyzed the EU institutional response to ethical and regulatory challenges of AI in addressing discriminatory bias, finding that the regulatory architecture is more developed than in other jurisdictions but still requires refinement for social media contexts [19]. Social media recommendation systems are not clearly classified as high-risk under the current framework, but pre-deployment bias assessment has been established as a regulatory norm likely to extend further [20].

The United States has no federal legislation targeting algorithmic gender bias specifically. The FTC has authority over unfair or deceptive practices and has used it in some algorithmic contexts, and state-level laws in California and Illinois impose transparency requirements on automated decision-making. China's Provisions on the Management of Algorithmic Recommendations, effective March 2022, require platforms to disclose recommendation criteria and allow users to adjust their settings, but the provisions focus on content moderation and political speech rather than gender bias [21]. All three systems have converged on transparency and audit requirements as primary tools. They diverge significantly on what substantive non-discrimination standards should actually demand.

DOI: 10.54254/2753-7048/2026.34114

4.3. Courts and proportionality

Courts applying the EU proportionality principle, which requires that restrictions on fundamental rights be appropriate, necessary, and proportionate to a legitimate aim, provide a framework for evaluating which regulatory interventions can be defended. Requiring platforms to audit recommendation systems for gender bias does not restrict expression at all; it only requires an assessment of whether content is being distributed in stereotypically gendered patterns. Requiring platforms to offer a non-personalized recommendation option is a modest obligation that the DSA already imposes [17]. Both interventions pass proportionality analysis without serious difficulty. The harder cases involve more substantive requirements, such as mandatory diversity parameters, where the cost to platform design autonomy is larger and the evidence of effectiveness is less settled.

5. A regulatory proposal

5.1. Mandatory auditing

The most direct intervention is requiring platforms to commission regular, independent audits of their recommendation systems for gender bias. Audits would assess whether the system distributes content in patterns correlated with inferred gender in ways that reflect stereotypical assumptions; whether the gender classification mechanisms in use are accurate, consensual, and contestable; and whether training data contains biases being reproduced in recommendation outputs. Noble has argued that algorithmic systems must be understood as exercises of power that require public accountability rather than private self-regulation [22]. A clearer legislative approach would classify recommendation algorithms above a defined user threshold as high-risk systems requiring mandatory auditing regardless of domain. Audit results would be submitted to a regulatory authority and summarized publicly. The neutrality defense becomes much harder to sustain when platforms are required to produce evidence rather than assertions.

5.2. Gender-inclusive design requirements

Auditing identifies problems in existing systems. A complementary requirement would shape the design of future ones. Legal frameworks should require developers to incorporate gender-inclusive considerations into recommendation architecture from the beginning, not as an afterthought. Practically, this could mean diversity parameters preventing any single content category from dominating a user's feed beyond a defined proportion. It could also mean requiring explicit, informed user consent before a gender label is assigned through behavioral inference, and giving users a clear mechanism to correct or reject the classification. Neither requirement removes content from platforms. Both push back against the engagement-maximizing logic that generates gender-typical recommendation patterns, and both give users more genuine control over their experience than current designs permit.

5.3. Transparency that is actually useful

The third component addresses user awareness. The EU DSA already requires platforms to explain to users how their recommendation systems function and to offer non-personalized alternatives [17]. In practice, the disclosures platforms produce under this requirement are typically too technical and too abstract to be genuinely informative. A more useful version of the requirement would oblige platforms to give users regular, readable summaries of their recommendation profile: what the

DOI: 10.54254/2753-7048/2026.34114

system has inferred about them, which content categories have been prioritized, and what steps are available to adjust. This does not resolve the structural problem, but it gives users a basis for informed engagement with the systems that shape their information environments, which is a precondition for any kind of meaningful individual agency.

6. Conclusion

The three mechanisms examined in this paper, echo chambers, automated gender classification, and historical data feedback loops, share a common feature: none of them requires any deliberate intent to discriminate in order to produce discriminatory outcomes. That is part of what makes them difficult to address. Existing anti-discrimination law is built around the concept of a decision, made by an actor, that can be evaluated against a standard. None of those elements maps cleanly onto algorithmic processes that generate biased outputs through the compounding logic of statistical inference applied to historical data.

The regulatory proposals offered here are designed to work within this constraint rather than against it. Mandatory auditing, gender-inclusive design requirements, and meaningful transparency obligations do not require characterizing any platform decision as discriminatory. They create structures for external accountability, design-stage correction, and user awareness that can function even where the harm is diffuse and intent is absent. Each proposal builds on legal tools that already exist in some form. The EU AI Act and DSA provide the closest analogues, though both require extension to cover social media recommendation specifically.

There is a broader question lurking behind the specific regulatory proposals. When a technology determines what gender representations billions of people encounter in their daily information environments, it is exercising a kind of social power. The question of whether that power is subject to governance is not really a technical question. It is a political and legal one, and the answer that law has generally given to comparable exercises of social power in other contexts is yes. There is no obvious reason algorithmic recommendation should be different. Future research should examine how the proposals here perform when implemented, whether similar frameworks can address bias along dimensions other than gender, and how regulatory coordination between jurisdictions might work given that platforms operate across borders with little regard for where their users are located.

References

[1] Global Witness. (2023). Facebook's algorithms discriminate by gender and race in job advertising [Research report].
[2] European Parliament. (2023). The impact of the use of social media on women and girls (Study PE 743.341). Policy Department for Citizens' Rights and Constitutional Affairs.
[3] Ellemers, N. (2018). Gender stereotypes. Annual Review of Psychology, 69, 275-298.
[4] Jamieson, K. H., and Capella, J. N. (2008). Echo chamber: Rush Limbaugh and the conservative media establishment. Oxford University Press.
[5] Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.
[6] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. Proceedings of the National Academy of Sciences of the United States of America, 118(9), e2023301118. https: //doi.org/10.1073/pnas.2023301118
[7] Ahmmad, M., Shahzad, K., Iqbal, A., and Latif, M. (2025). Trap of social media algorithms: A systematic review of research on filter bubbles, echo chambers, and their impact on youth. Societies, 15(11), 301. https: //doi.org/10.3390/soc15110301
[8] Fosch-Villaronga, E., Mut-Piña, A., Verhoef, T., Poulsen, A., Søraa, R. A., and Custers, B. H. M. (2026). Misgendering algorithms: Insights from a cross-sectional survey on algorithmic gender classification in social media. Technology in Society, 84, 103110. https: //doi.org/10.1016/j.techsoc.2026.103110

DOI: 10.54254/2753-7048/2026.34114

[9] Fosch-Villaronga, E., Poulsen, A., Søraa, R. A., and Custers, B. H. M. (2021). Gendering algorithms in social media. ACM SIGKDD Explorations Newsletter, 23(1), 24-32. https: //doi.org/10.1145/3468507.3468512
[10] Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., and Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. Psychological Science, 32(2), 218-240. https: //doi.org/10.1177/0956797620963618
[11] Rathee, S., Banker, S., Mishra, A., and Mishra, H. (2023). Algorithms propagate gender bias in the marketplace with consumers' cooperation. Journal of Consumer Psychology, 33(4), 621-631. https: //doi.org/10.1002/jcpy.1351
[12] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.
[13] Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.
[14] Wachter, S., and Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Columbia Business Law Review, 2019(2), 494-620.
[15] European Parliament Research Service. (2025). Algorithmic discrimination under the AI Act and the GDPR (EPRS Briefing PE 769.509).
[16] O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.
[17] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services (Digital Services Act). Official Journal of the European Union, L 277, 1-102.
[18] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (AI Act). Official Journal of the European Union, L 2024/1689.
[19] Cerezo-Martínez, P., Nicolás-Sánchez, A., and Castro-Toledo, F. J. (2024). Analyzing the European institutional response to ethical and regulatory challenges of artificial intelligence in addressing discriminatory bias. Frontiers in Artificial Intelligence, 7, article 1393259. https: //doi.org/10.3389/frai.2024.1393259
[20] Lütz, F. (2024). The AI Act, gender equality and non-discrimination: What role for the AI office? ERA Forum, 25, 79-95. https: //doi.org/10.1007/s12027-024-00785-w
[21] Provisions on the Management of Algorithmic Recommendations. Cyberspace Administration of China. Effective 1 March 2022.
[22] Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.