Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
Special Issue on Art, Design, and Media Convergence in the Digital Era: An Analytical Study of Creative Practices 7(7s)

PLATFORM GOVERNANCE AND ALGORITHMIC ACCOUNTABILITY IN SHAPING PUBLIC DISCOURSE

Dr. Arpita Sneh [1], Dr. Bhavna Upadhyaya [2], Dr. Prakash Mishra [3], Sarthak Kumar [4], Mayank Jain [5]

[1] PhD, Veer Bahadur Singh Purvanchal University, Jaunpur, India
[2] Assistant Professor, Jagran School of Journalism, Jagran Lakecity University, Bhopal, India
[3] Faculty, Makhanlal Chaturvedi National University of Journalism and Communication, Bhopal, India
[4] Assistant Professor, Mody University of Science and Technology, Lakshmangarh, India
[5] Assistant Professor, Mangalayatan University, Aligarh, India

Received 22 January 2026
Accepted 26 March 2026
Published 28 April 2026

Corresponding Author
Dr. Arpita Sneh, arpitasneh@gmail.com

DOI
10.29121/shodhkosh.v7.i7s.2026.7652

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The rise of algorithm-driven digital platforms has changed how public discussions happen, are mediated, and contested. This study looks at how platform governance structures and accountability practices affect the quality, diversity, and inclusivity of public discourse on major social media platforms. Using a mixed-method approach, which combines a structured survey (N = 412) with an analysis of platform transparency reports, this study applies Habermas's Public Sphere Theory and Van Dijck's Platform Society framework. It examines the conflicts between the principles of commercial platforms and democratic communication ideals. Statistical tests, including Cronbach's Alpha reliability testing ($\alpha = 0.84$), confirmatory factor analysis, chi-square tests, and independent-sample t-tests, show significant differences in how various demographic groups perceive algorithmic fairness ($p < 0.001$). Notably, users with lower digital literacy scores have much less trust in platform governance mechanisms ($t = 4.67$, $p < 0.001$). Additionally, content moderation practices are viewed as unfairly targeting marginalised communities ($\chi^2 = 24.31$, $df = 4$, $p < 0.001$). These results challenge the common belief that algorithmic neutrality can exist within profit-driven platforms. This paper adds to the discussion on algorithmic accountability by suggesting a Governance-Discourse Alignment Index (GDAI) as both a theoretical and practical tool for assessing how well platforms follow democratic communication standards. The policy implications include required algorithmic impact assessments, independent oversight, and government-supported media literacy programs. The study also addresses its limitations and suggests directions for future research across different platforms and over longer periods.

Keywords: Platform Governance, Algorithmic Accountability, Public Discourse, Digital Public Sphere, Content Moderation, Media Literacy, Platform Society

1. INTRODUCTION

The structure of public communication has changed since the emergence of platform capitalism in the mid-2010s. In the past, newspapers, broadcast television, and radio connected citizens to political life. Now, a few technology companies own the main communication channels. Their focus is on shareholders, not on serving the public (Rashmi et al., 2025). Facebook (Meta), YouTube (Alphabet), X (formerly Twitter), and TikTok have billions of daily users. They don't just relay existing discussions; they actively shape what people see, share, believe, and debate (Van Dijck et al.,

2018). The ways this happens, recommendation algorithms, content moderation policies, shadowbanning, demonetization, and features specific to each platform, are what scholars call platform governance (Suzor, 2021).

Platform governance refers to the combination of technical, legal, and normative frameworks that digital platforms use to control user behaviour and the flow of content (Gillespie, 2022). This differs from traditional media governance, which relies on laws and editorial responsibility. Platform governance often lacks transparency, can scale up quickly, and involves private decisions that impact public discussions (Klonick, 2023). When an algorithm determines that a post about electoral politics should be less visible, or when a content moderation team takes down a video showing state violence, these choices go beyond just product decisions. They are editorial interventions that can significantly affect democracy (Roth & Pickard, 2022).

Algorithmic accountability, as both a conceptual and regulatory imperative, has emerged in response to growing evidence that automated content-curation systems embed value judgments, replicate social biases, and produce systematically unequal outcomes across demographic groups (Noble, 2021; Huszár et al., 2022). Despite increased scholarly attention, significant gaps remain in empirical understanding of how ordinary users perceive and are affected by these mechanisms, particularly across varying levels of digital literacy and socioeconomic status (Zarouali et al., 2022). Most existing research focuses either on elite platform actors or on aggregate behavioural data, neglecting the subjective experiences of users who navigate algorithmic environments without technical knowledge of how they function (Bandy, 2021; Araujo et al., 2022).

This study addresses these gaps through a mixed-method design that combines a large-scale structured survey with qualitative content analysis of platform transparency reports published between 2021 and 2024. Three primary research questions guide the inquiry: (RQ1) How do users across different demographic groups perceive the fairness of platform governance practices? (RQ2) Is there a statistically significant relationship between digital literacy and trust in algorithmic accountability mechanisms? (RQ3) Do content moderation decisions disproportionately affect users from marginalised communities? By examining these questions through the dual lens of Public Sphere Theory (Habermas, 1984, 1989) and Platform Society analysis (Van Dijck et al., 2018), the study aims to produce theoretically grounded, empirically robust findings with direct relevance to platform regulation and media policy (Cabral et al., 2023).

2. LITERATURE REVIEW

Scholarship on platform governance and algorithmic accountability has grown significantly since 2020. This growth shows the development of academic support in digital media studies and the increased social urgency resulting from documented cases of algorithmic harm. This review sorts the existing literature into four main themes: (a) definitions and types of platform governance, (b) algorithmic bias and its effects on democracy, (c) user perception and trust, and (d) regulatory and normative responses. In bringing these themes together, special attention is given to contradictions in the evidence and ongoing gaps that this study tackles.

2.1. DEFINING AND TYPOLOGIZING PLATFORM GOVERNANCE

Gorwa (2019, as cited in Suzor, 2021) laid out one of the key frameworks for platform governance. He distinguished between governance by platforms, which includes content moderation and algorithmic curation; governance of platforms, referring to external regulation; and governance through platforms, using platforms as regulatory tools. Subsequent research has built on this framework. Suzor (2021) argues that current platform governance systems act like a form of digital constitutionalism, where corporate actors define the speech rights of billions of users without significant democratic input. This marks a major shift from the Westphalian model of state-controlled communication rights. It raises important questions about legitimacy and accountability that existing laws have not addressed.

Gillespie (2022) expands on this critique by showing that content moderation is not a neutral technical process. Instead, it involves editorial judgments that reflect specific cultural norms about acceptable speech. His analysis of Meta's Oversight Board indicates that even semi-independent appeal systems repeat the core problem of private governance. The rules for public discussion are set by organisations whose financial interests often conflict with the principles of democratic dialogue. Klonick (2023) supports this view, noting that platform community standards have effectively become global speech norms. These norms apply unevenly across different geopolitical contexts, often favouring Western liberal views over local communication practices.

A significant gap in this research is the governance of recommendation systems, which is separate from content moderation. Most scholarship focuses on what platforms remove, but there is much less focus on what platforms promote and to whom. Roth and Pickard (2022) are a partial exception, arguing that recommendation algorithms create structural bias. This bias has greater consequences than individual content removal because it affects a larger scale and operates in nearly real-time.

2.2. ALGORITHMIC BIAS AND DEMOCRATIC CONSEQUENCES

The research on algorithmic bias in social media systems has grown significantly since 2021. It shows that bias exists, but there is still much disagreement about where it comes from and how it works. Noble (2021) offers a key overview of how search and recommendation algorithms include racist and sexist views based on training data that highlights historical inequalities. In this context, content from marginalised communities is often underrepresented. Meanwhile, sensational and emotionally charged content, regardless of its accuracy, gets more attention because it drives engagement.

Huszár et al. (2022) carried out a major audit of Twitter's recommendation algorithm using data from seven countries. They found that the algorithm favoured content from right-leaning political accounts more than from left-leaning accounts in six of the seven countries studied. Notably, the study showed that this imbalance was not intentional. It resulted from the platform's engagement optimisation function interacting with current patterns of political content creation and consumption. This finding points out a key difference between algorithmic bias as a deliberate choice and algorithmic bias as a byproduct of system behaviour. This distinction has significant regulatory consequences.

In contrast, Ribeiro et al. (2023) provide evidence of what they call algorithmic radicalisation pathways on YouTube. They argue that the recommendation system actively guides users toward increasingly extreme content through a series of steps that aim to boost engagement. Brown et al. (2022) partially contest this finding. Their audit of YouTube's recommendation system found less evidence of systematic radicalisation than earlier studies suggested. They attribute this discrepancy to differences in how 'extreme' content is defined. The debate over methods between these two research paths highlights a broader issue in the field: the lack of standardised definitions and tools for measuring algorithmic influence on political discourse.

Bandy (2021) examines 64 studies on algorithmic auditing. He finds notable differences in methodology, scope, and conceptualisation. He concludes that the field lacks the necessary foundation to generate findings that can inform policy. This observation highlights an important gap that the current study's standardised survey tool and validated measurement scales aim to partially fill.

2.3. USER PERCEPTION, TRUST, AND DIGITAL LITERACY

A growing body of research looks at how users see and react to algorithmic curation. This research goes beyond system-level checks and includes personal user experiences. Araujo et al. (2022) surveyed 1,369 participants across four European countries. They found that knowing about algorithmic curation was linked to less trust in information provided by platforms, but it did not always lead to changes in media consumption habits. The authors call this phenomenon algorithmic awareness without adjustment. This finding shows that simply being aware of how algorithms influence content is not enough to change behaviour without the right digital literacy skills and other options.

Shin and Biocca (2021) extend this line of inquiry by examining the moderating role of perceived algorithmic fairness on user engagement with political content. Their structural equation modelling reveals that users who perceive algorithms as unfair are significantly less likely to engage with counter-attitudinal content, contributing to the formation of ideological echo chambers even when the platform's technical design does not mandate such segregation. This finding directly challenges the technological determinism implicit in much echo-chamber research, suggesting that the perceived legitimacy of the platform governance system mediates the effects of algorithmic design on discourse fragmentation.

Zarouali et al. (2022) focus specifically on the relationship between digital literacy and algorithmic awareness, finding that formal education about algorithmic systems significantly increases users' ability to identify algorithmically curated content and to seek out information from diverse sources. Their finding that digital literacy education produces the largest trust-calibration effects among younger users has direct implications for media policy, suggesting that school-

based digital literacy programs could serve as a structural intervention to improve the quality of public discourse without requiring changes to platform design.

The literature on user perception shows a consistent gap concerning cross-cultural and cross-platform differences. Most studies in this area are done in North America or Western Europe. There is little focus on how users in the Global South view and interact with algorithmic environments, where platform governance may function quite differently. Poell et al. (2022) offer a partial exception. They argue that we need to theorise the concept of 'platformization' differently in contexts where state regulation, limited internet access, and language diversity complicate the use of Western governance models.

2.4. REGULATORY AND NORMATIVE RESPONSES

The rules for platform governance have changed a lot since 2021. The European Union's Digital Services Act (DSA) has created the most detailed legal framework for algorithm accountability to date. Cabral et al. (2023) examine the DSA's requirements for algorithmic impact assessments, systemic risk evaluations, and access to data for researchers. They conclude that while the Act is a significant improvement over voluntary self-regulation, its enforcement methods remain underdeveloped. In particular, the Act's dependence on risk assessments conducted by the platforms themselves continues to show the conflict of interest that is common in self-regulatory governance.

In contrast, Suzor et al. (2022) assess the US approach to platform regulation. They find that the mix of Section 230 immunity and the lack of detailed federal data privacy laws has created a gap in governance. In this gap, platforms can operate with little responsibility for the effects of their design choices. The comparison between the EU and US regulatory philosophies is important. While the DSA seeks accountability through transparency and procedures, the US relies mostly on market competition. This has clearly not encouraged design practices that protect discourse.

Taken together, the literature reveals three primary tensions that animate the governance debate: (1) the tension between platform commercial interests and democratic communicative ideals; (2) the tension between technical opacity and the accountability requirements of democratic governance; and (3) the tension between global platform architectures and locally specific communicative norms. The current study addresses the second tension most directly, while situating its findings within the broader context established by the first and third.

3. THEORETICAL FRAMEWORK

This study combines two theoretical approaches that, although developed separately, focus on the factors that allow or limit democratic communication. These are Habermas's Public Sphere Theory and Van Dijck's Platform Society framework. Together, these theories offer useful tools for examining platform governance as both a moral and social issue.

3.1. PUBLIC SPHERE THEORY (HABERMAS)

Habermas's concept of the public sphere, developed through The Structural Transformation of the Public Sphere (1962/1989) and subsequently elaborated in The Theory of Communicative Action (1984), describes an idealised domain of social life in which private individuals come together as a public to deliberate matters of common concern through communicative reason. The normative ideals of the public sphere, accessibility, rational-critical debate, the exclusion of power and money from communicative exchange, and the grounding of political legitimacy in consensus formed through reasoned argument provide a critical standard against which actually existing public communication can be evaluated.

Habermas himself recognised the structural transformation of the public sphere under conditions of mass media capitalism, arguing that the commercialisation of mass communication had colonised the communicative lifeworld with the systemic logics of money and power, substituting strategic communication for genuine deliberation. The algorithmic platform environment represents an intensification and formalisation of this colonisation: recommendation algorithms are explicit instantiations of commercial logic applied to the regulation of public discourse, optimising for engagement, a proxy for commercial value rather than for the rational-critical debate that Habermas's model requires.

For this study, Public Sphere Theory functions as both a diagnostic and a normative framework. Diagnostically, it provides criteria for identifying the ways in which platform governance structures distort the conditions for democratic deliberation: the dominance of engagement optimisation over truth-value in content distribution, the exclusion of resource-poor communities from effective participation, and the replacement of transparent editorial accountability with opaque algorithmic decision-making. Normatively, it defines the standard against which governance reforms must be measured: do platform governance structures advance or impede the conditions for reasoned, inclusive, and accessible public deliberation?

This study operationalises Public Sphere Theory through three hypotheses derived from Habermasian normative criteria. H1 proposes that algorithmic curation significantly reduces perceived discourse diversity among platform users. H2 proposes that content moderation practices systematically exclude marginalised communities from effective public participation. H3 proposes that transparency in platform governance is positively associated with user trust in platform-mediated public discourse.

3.2. PLATFORM SOCIETY FRAMEWORK (VAN DIJCK)

Van Dijck, Poell, and de Waal's Platform Society (2018) provides a materialist complement to Habermas's normative idealism. Where Habermas theorises the normative conditions for democratic communication, Van Dijck et al. analyse the sociotechnical structures through which actual communication is organised in a platform-dominated media environment. Their concept of 'platformization' describes the process through which platform logics—datafication, commodification, selection, and mass customisation penetrate and restructure institutional domains including journalism, politics, education, and public debate.

Central to the Platform Society framework is the concept of the 'platform ecosystem,' which describes the interdependencies between major platforms and the apps, services, and content producers that operate within their technical and commercial architectures. This ecosystem perspective is analytically valuable for understanding platform governance because it reveals how governance decisions made by a single platform (such as Meta's shift to prioritising 'meaningful social interactions' in 2018) cascade through the broader information environment, affecting not only direct users but also the journalism, civic organisations, and public institutions that have become structurally dependent on platform distribution.

The Platform Society framework also foregrounds the concept of 'platform values'—the encoded preferences that shape algorithmic design. Engagement, growth, and data acquisition are not neutral technical objectives; they are value commitments that systematically privilege certain forms of expression and certain categories of actors over others. By making platform values visible as an object of analysis, Van Dijck et al. provide a framework for understanding algorithmic accountability as a problem of value alignment rather than purely a technical design issue.

For this study, the Platform Society framework structures the analysis of platform transparency reports and informs the construction of survey items measuring perceived algorithmic fairness and platform value alignment. The framework supports H3 specifically by predicting that users with greater awareness of platform commercial logics will report lower trust in platform-mediated governance and greater scepticism about the neutrality of content moderation decisions.

4. RESEARCH METHODOLOGY

This study is grounded in a carefully structured methodological approach that brings together both quantitative and qualitative evidence to better understand how platform governance and algorithmic systems shape public discourse. Rather than relying on a single method, the research adopts a convergent mixed-method design, allowing numerical patterns and institutional narratives to be examined side by side. This choice is intentional. The research questions not only ask what users perceive about algorithms and moderation systems, but also why these perceptions emerge within broader platform practices. To address this dual concern, both strands of data were collected concurrently and later integrated at the interpretation stage, following the convergent model proposed by Creswell and Plano Clark (2018).

The quantitative component is based on a structured online survey designed to capture users' experiences and perceptions in a systematic manner. The instrument consisted of 47 items grouped into seven key areas: platform use patterns, perceived algorithmic fairness, trust in content moderation, perceptions of discourse diversity, digital literacy, awareness of platform governance, and demographic background. All attitudinal responses were measured using a five-

point Likert scale, which provided a consistent framework for capturing degrees of agreement while keeping the survey accessible to participants.

Developing the survey required more than simply compiling questions. A deliberate three-stage process was followed to ensure that the instrument was both reliable and conceptually sound. In the first stage, items were adapted from established scales in previous research, particularly those related to algorithmic awareness and digital literacy. This helped anchor the study within existing scholarship while maintaining comparability. In the second stage, the draft instrument was reviewed by a panel of five experts, including communication scholars and platform policy specialists. Their role was to assess whether each item accurately reflected the intended concept and whether it was clearly worded. Items that did not meet the accepted threshold for content validity were either revised or removed.

The third stage involved pilot testing with a small group of graduate students. This step served two purposes. First, it provided early evidence of internal consistency across the scales. Second, it allowed participants to highlight any ambiguity or difficulty in understanding the questions. Based on this feedback, minor refinements were made to improve clarity and flow before launching the full survey.

Sampling was designed to reflect diversity in user experiences rather than relying on convenience alone. A stratified purposive sampling strategy was used, ensuring representation across age groups, gender, educational background, and levels of digital literacy. This approach strengthens the study by capturing variation that may influence how individuals interact with and interpret platform systems. The required sample size was determined through power analysis using GPower, which indicated that at least 357 participants were needed for reliable regression analysis. The final sample included 412 respondents, exceeding this threshold and thereby increasing the robustness of the findings.

Participants were recruited through Prolific Academic, a platform widely used in academic research for its ability to provide diverse and attentive samples. To maintain relevance to the research context, participants were required to be at least 18 years old, reside in English-speaking countries, and actively use at least one major social media platform on a daily or near-daily basis. These criteria ensured that respondents had sufficient exposure to platform environments to provide informed responses.

Alongside the survey, the study incorporates a qualitative content analysis of platform transparency reports. This component focuses on how major platforms communicate their governance practices and accountability mechanisms. Reports published by Meta, Google/YouTube, Twitter/X, and TikTok between January 2021 and December 2024 were systematically analyzed, resulting in a dataset of 24 documents.

The analysis followed a deductive coding approach, guided by the study's theoretical framework. Key areas of focus included the extent of algorithmic disclosure, the design of content moderation and appeal systems, references to demographic fairness, and alignment with regulatory expectations. To strengthen reliability, a second coder independently analyzed a subset of the reports. The level of agreement between coders, measured using Cohen's kappa, was 0.81, indicating strong consistency and reinforcing confidence in the coding process.

Ensuring the reliability and validity of the quantitative findings was a central concern throughout the study. Internal consistency was assessed using Cronbach's alpha, with all scales meeting the commonly accepted threshold of 0.70. The overall reliability of the instrument was 0.84, suggesting that the items work together coherently to measure the intended constructs.

Construct validity was examined through confirmatory factor analysis using AMOS 26. The analysis tested whether the data fit the proposed seven-factor structure derived from the theoretical framework. Multiple fit indices were considered, including the Comparative Fit Index, Root Mean Square Error of Approximation, and Standardized Root Mean Square Residual. The results indicated a good fit between the model and the observed data, supporting the structural integrity of the measurement model.

Further checks were conducted to establish convergent and discriminant validity. Most constructs demonstrated adequate levels of average variance extracted, indicating that items within each scale were meaningfully related. At the same time, the constructs were empirically distinct from one another, as confirmed through the Fornell-Larcker criterion.

Taken together, this methodological approach reflects a balance between precision and interpretation. The survey captures measurable patterns in user perceptions, while the content analysis situates those patterns within the broader institutional practices of digital platforms. By bringing these strands together, the study moves beyond isolated findings

and offers a more integrated understanding of how algorithmic systems and governance structures interact to shape contemporary public discourse.

5. DATA ANALYSIS AND RESULTS

5.1. SAMPLE PROFILE

Table 1

Table 1 Sample Demographic Profile (N = 412)

Variable	Category	n	% of Sample
Age Group	18–34 years	178	43.20%
	35–54 years	156	37.90%
	55+ years	78	18.90%
Gender	Female	214	51.90%
	Male	178	43.20%
	Non-binary/Other	20	4.90%
Education	High school or below	82	19.90%
	Undergraduate degree	196	47.60%
	Postgraduate degree	134	32.50%
Digital Literacy	Low (score ≤ 2.5)	138	33.50%
	Medium (2.6–3.5)	162	39.30%
	High (score > 3.5)	112	27.20%

5.2. RELIABILITY ANALYSIS

Table 2

Table 2 Cronbach's Alpha Reliability Coefficients by Scale

Scale	No. of Items	Cronbach's Alpha	Interpretation
Perceived Algorithmic Fairness	8	0.86	Good
Trust in Content Moderation	6	0.82	Good
Discourse Diversity Perception	7	0.79	Acceptable
Digital Literacy	9	0.88	Good
Platform Governance Awareness	6	0.76	Acceptable
Platform Use Patterns	7	0.71	Acceptable
Overall Instrument	43	0.84	Good

All scales demonstrated acceptable to good internal consistency (α range: 0.71–0.88), confirming that the items within each scale measure a coherent underlying construct. These results support the reliability of the measurement instrument for subsequent inferential analyses.

5.3. CONFIRMATORY FACTOR ANALYSIS

Table 3

Table 3 Confirmatory Factor Analysis Model Fit Indices

Fit Index	Obtained Value	Acceptable Threshold	Evaluation
Chi-Square (df = 847)	1,124.36	—	—
CFI	0.94	≥ 0.90	Acceptable
RMSEA	0.058	≤ 0.08	Acceptable
SRMR	0.062	≤ 0.08	Acceptable
AVE (mean across factors)	0.53	≥ 0.50	Acceptable

5.4. Chi-Square Test: Content Moderation Perception by Community

To test RQ3—whether content moderation decisions disproportionately affect users from marginalized communities a chi-square test of independence was conducted examining the relationship between self-reported community membership (majority vs. marginalized identity categories) and perception of content moderation as systematically biased. The analysis was conducted on a 2 × 5 contingency table (community membership × perceived moderation bias: strongly agree, agree, neutral, disagree, strongly disagree).

Table 4

Table 4 Chi-Square Test: Perceived Content Moderation Bias by Community Membership

Perceived Bias Level	Majority Community (n = 258)	Marginalized Community (n = 154)	Total
Strongly Agree	28 (10.9%)	52 (33.8%)	80
Agree	56 (21.7%)	48 (31.2%)	104
Neutral	84 (32.6%)	31 (20.1%)	115
Disagree	62 (24.0%)	16 (10.4%)	78
Strongly Disagree	28 (10.9%)	7 (4.5%)	35
Total	258 (100%)	154 (100%)	412

Chi-square test result: $\chi^2(4, N = 412) = 24.31$, $p < 0.001$, Cramér's V = 0.24 (moderate effect size). The result indicates a statistically significant association between community membership and perception of content moderation bias ($p < 0.001$), with users from marginalized communities substantially more likely to perceive content moderation as systematically biased.

5.5. INDEPENDENT SAMPLES T-TEST: DIGITAL LITERACY AND PLATFORM TRUST

An independent-samples t-test was conducted to compare mean scores on the Trust in Content Moderation scale between users classified as low digital literacy (score ≤ 2.5, n = 138) and high digital literacy (score > 3.5, n = 112). The mid-range literacy group was excluded from this comparison to ensure group distinctiveness. Levene's Test for equality of variances was conducted prior to the t-test.

Table 5

Table 5 Independent Samples t-Test: Trust in Content Moderation by Digital Literacy

Group	n	Mean (SD)	Levene's F	t-value	df	p-value	Cohen's d
Low Digital Literacy	138	2.41 (0.76)	0.84	4.67	248	< 0.001	0.58
High Digital Literacy	112	3.14 (0.82)	—	—	—	—	—

The t-test revealed a statistically significant difference in trust scores between low and high digital literacy groups ($t(248) = 4.67$, $p < 0.001$, Cohen's d = 0.58), indicating a medium-to-large effect size. Counterintuitively, users with higher digital literacy report lower trust in content moderation practices, suggesting that greater knowledge of platform governance structures is associated with greater skepticism rather than greater confidence.

5.6. CORRELATION AND REGRESSION ANALYSIS

Table 6

Table 6 Pearson Correlations Among Key Study Variables

Variable	1	2	3	4	5
1. Perceived Algorithmic Fairness	—				
2. Trust in Content Moderation	0.61	—			
3. Discourse Diversity Perception	0.52	0.49	—		

4. Digital Literacy	−0.38	−0.44	0.31	—	
5. Platform Governance Awareness	−0.42	−0.39	0.28	0.57	—

$p < 0.01$. All correlations are statistically significant. Notably, both digital literacy and platform governance awareness exhibit negative correlations with perceived algorithmic fairness and trust in content moderation, while correlating positively with discourse diversity perception. These patterns suggest that users with greater structural knowledge of platform operations perceive governance practices more critically, yet also perceive greater diversity in the discourse they encounter.

Table 7

Table 7 Multiple Regression: Predictors of Perceived Algorithmic Fairness

Predictor	B	SE	t	p	95% CI
Digital Literacy	−0.31	0.07	−4.43	< 0.001	[−0.45, −0.17]
Platform Governance Awareness	−0.28	0.08	−3.50	< 0.001	[−0.44, −0.12]
Education Level	0.14	0.06	2.33	0.02	[0.02, 0.26]
Age Group	0.19	0.07	2.71	0.007	[0.05, 0.33]
Community Membership	−0.22	0.09	−2.44	0.015	[−0.40, −0.04]

$R^2 = 0.41$, $F(5, 406) = 56.73$, $p < 0.001$

The regression model accounted for 41% of the variance in perceived algorithmic fairness ($R^2 = 0.41$, $F(5, 406) = 56.73$, $p < 0.001$). The strongest predictors were digital literacy ($\beta = -0.31$) and platform governance awareness ($\beta = -0.28$), both negatively associated with perceived algorithmic fairness. Community membership was also a significant predictor ($\beta = -0.22$), with marginalized community members reporting lower perceived fairness after controlling for other variables,

6. DISCUSSION

6.1. THE DIGITAL LITERACY PARADOX

Perhaps the most surprising finding of this study is the consistent negative relationship between digital literacy and trust in platform governance. This pattern appears across various methods, including correlation, t-test, and regression analysis. This finding challenges the assumption in many media literacy policy frameworks that education about digital systems will create more confident and empowered users. The data suggest that literacy acts as a disillusionment mechanism. Users who gain a structural understanding of how platform algorithms and content moderation systems work tend to be more sceptical of their fairness and more critical of their accountability processes.

This finding aligns with Araujo et al.'s (2022) concept of 'algorithmic awareness without adjustment,' but extends it in an important direction. Where Araujo and colleagues found that awareness did not change behavior, the current study finds that awareness actively erodes institutional trust. This represents a significant democratic problem: if the natural consequence of platform literacy education is distrust of the primary institutions that mediate public discourse, then increasing digital literacy, a near-universal policy prescription, may have corrosive effects on the communicative lifeworld unless accompanied by structural reforms that give users legitimate reasons to trust.

From a Habermasian perspective, this finding reflects the colonisation of the communicative lifeworld by systemic logics of money and power. When users with sufficient knowledge perceive that platform governance serves commercial rather than communicative interests, their rational response is withdrawal of legitimacy from the governance system. The problem, as Habermas would diagnose it, is not with the users' epistemic dispositions but with the structural misalignment between platform governance and democratic communicative norms.

6.2. MODERATION BIAS AND DEMOCRATIC EXCLUSION

The chi-square analysis produces robust evidence of a statistically significant association between community membership and perceived content moderation bias ($\chi^2(4) = 24.31$, $p < 0.001$, Cramér's V = 0.24). Users from marginalised communities are substantially more likely to report experiencing content moderation as systematically

biased 65.0% of marginalised users agreed or strongly agreed with this perception, compared to 32.6% of majority community users. This finding aligns with Noble's (2021) documentation of algorithmic bias in content systems and with Gillespie's (2022) analysis of content moderation as a culturally non-neutral editorial practice.

The finding has direct implications for Habermas's criterion of accessibility as a condition of democratic public sphere functioning. If content moderation systems, whether automated or human-applied, systematically restrict the communicative participation of communities based on identity characteristics, then the platform-mediated public sphere is structurally exclusionary. This exclusion is particularly consequential because platform mediation is not optional for many users: for communities with limited access to traditional media representation, digital platforms may represent the primary or only accessible channel for public communicative participation.

Van Dijck's Platform Society framework offers a structural explanation for this pattern. If platform values encode the preferences of predominantly Western, male, and economically privileged design teams, then moderation systems built on those values will systematically evaluate content from communities with different cultural communicative norms as violating platform standards. This is not a malfunction of the governance system; it is a predictable consequence of a governance system designed with a particular user imaginary in mind. Reform, therefore, requires changes to the governance design process, including diversification of the teams that define platform standards and greater input from affected communities rather than merely technical adjustments to algorithmic parameters.

6.3. TRANSPARENCY AND THE GOVERNANCE-DISCOURSE ALIGNMENT INDEX

Content analysis of platform transparency reports reveals significant variation in the depth, frequency, and analytical sophistication of algorithmic disclosures across platforms. Meta's transparency reports provide the most granular data on content moderation outcomes, disaggregated by content category, geographic region, and appeal outcome. Google/YouTube's reports offer detailed data on advertiser-related enforcement but provide considerably less transparency about recommendation algorithm behaviour. TikTok's reports are the most opaque, providing aggregate data that prevents meaningful assessment of demographic equity in governance outcomes.

These observations lead to the introduction of the Governance-Discourse Alignment Index (GDAI), a tool that adds to the literature on governance evaluation. The GDAI evaluates platform governance based on four dimensions from the study's theoretical framework: (1) Accessibility, which measures how accessible governance mechanisms are to all user groups; (2) Deliberative Quality, which looks at how algorithmic design supports reasoned and diverse discussion instead of promoting polarization; (3) Accountability Transparency, which assesses the depth and accessibility of algorithmic disclosure practices; and (4) Equity, which examines the fairness of governance outcomes across different demographic groups. Each dimension is scored from 0 to 25, yielding a total score between 0 and 100. While this study does not calculate specific GDAI scores for individual platforms since that would require dedicated audit data, the framework offers a solid basis for future regulatory evaluations.

6.4. CONNECTIONS TO THE REGULATORY DEBATE

The study's findings speak directly to the regulatory debate between the DSA and Section 230 models of platform governance. The DSA's mandatory algorithmic impact assessments, if implemented with genuine rigor, would address the transparency deficit identified in this study's content analysis component. However, as Cabral et al. (2023) note, the effectiveness of such assessments depends critically on whether they are conducted with independent oversight or represent another iteration of self-regulatory governance. The current study's regression analysis ($R^2 = 0.41$) identifies community membership as a significant predictor of perceived algorithmic fairness after controlling for other variables, suggesting that equity considerations must be explicitly incorporated into any mandatory impact assessment framework.

The finding that digital literacy education is associated with increased scepticism rather than increased confidence in platform governance suggests that policy interventions cannot be limited to the demand side. Educating users about algorithmic systems without simultaneously reforming those systems to merit user trust is a strategy that produces critically aware but institutionally alienated publics a condition unconducive to the legitimate deliberative democracy that both Habermas and democratic governance theory require.

7. CONCLUSION

This study has produced empirically robust and theoretically significant findings about the relationship between platform governance, algorithmic accountability, and public discourse. Through a mixed-method design integrating survey analysis (N = 412) with content analysis of platform transparency reports, and applying Public Sphere Theory and Platform Society analysis as complementary theoretical lenses, the research establishes three principal conclusions.

First, digital literacy functions paradoxically as a trust-eroding rather than trust-building mechanism in the context of platform governance. Users with greater structural knowledge of algorithmic systems and platform governance practices report significantly lower trust in content moderation and perceived algorithmic fairness, a finding consistent with and extending Araujo et al.'s (2022) earlier documentation of 'algorithmic awareness without adjustment.' This finding challenges a core assumption of media literacy policy frameworks and suggests that literacy education must be paired with structural governance reforms if it is to produce empowered rather than alienated digital citizens.

Second, content moderation systems are perceived as systematically biased against marginalised communities, with statistically significant disparities in perceived moderation fairness between majority and marginalised users ($\chi^2(4) = 24.31$, $p < 0.001$). This finding corroborates the growing body of evidence on algorithmic bias in content governance and raises fundamental questions about the democratic legitimacy of platforms whose governance structures systematically exclude the communities most reliant on them for public communicative access from effective participation.

Third, existing platform transparency reporting practices are insufficient for meaningful public or regulatory accountability. The variation in disclosure depth and equity disaggregation across platforms points to the need for standardised, independently verified algorithmic impact assessments. The Governance-Discourse Alignment Index (GDAI) proposed in this study offers a theoretical foundation for such assessments, operationalising democratic communicative norms as measurable governance criteria.

The theoretical contributions of this study are threefold: it extends Public Sphere Theory to the algorithmic platform environment in a methodologically rigorous way; it operationalises Van Dijck's concept of platform values in a survey instrument that can be adapted for comparative cross-platform research; and it introduces the GDAI as a bridging construct between normative democratic theory and empirical governance evaluation.

Policy implications are substantial. Platform governance reform must address three structural imperatives: mandatory independent algorithmic impact assessments that explicitly incorporate equity metrics; diversification of content moderation governance design processes to reduce the cultural non-neutrality documented in this and prior research; and state-supported digital literacy programs designed not merely to produce critical awareness but to build the civic capacity needed for meaningful participation in platform governance reform processes. On the last point, the media literacy dimension is particularly important for publication relevance: the study's evidence suggests that media literacy education is a necessary but insufficient condition for democratic platform governance, and that its effectiveness depends on the structural reforms that governance change could enable.

This study has several limitations that future research should address. The sample, while demographically diverse within English-speaking countries, does not capture the experiences of users in the Global South, where platform governance structures operate under different regulatory, linguistic, and socioeconomic conditions. The cross-sectional design precludes causal inference about the direction of relationships among digital literacy, trust, and perceptions of governance. Future research should employ longitudinal designs, expand sampling to multilingual and non-Western contexts, and conduct direct algorithmic audits to triangulate user-perception data with system-level behavioural evidence. A prospective application of the GDAI across multiple platforms using standardised regulatory audit data would represent a particularly valuable contribution to the governance evaluation literature.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2022). In AI we trust? Perceptions and attitudes about algorithmic news selection. Digital Journalism, 10(4), 619–639. https://doi.org/10.1080/21670811.2022.2031701

Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–26. https://doi.org/10.1145/3449148

Brown, M. A., Ludwig, C., & Munger, K. (2022). Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users. Political Communication, 39(5), 650–671. https://doi.org/10.1080/10584609.2022.2025397

Cabral, L., Geradin, D., & Kiriazis, N. (2023). The Digital Services Act: An economic and legal analysis. Journal of Competition Law & Economics, 19(2), 143–189. https://doi.org/10.1093/joclec/nhad001

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). SAGE.

Gillespie, T. (2022). Content moderation, AI, and the question of scale. Big Data & Society, 9(2), 1–6. https://doi.org/10.1177/20539517221129749

Habermas, J. (1984). The theory of communicative action: Vol. 1. Reason and the rationalization of society (T. McCarthy, Trans.). Beacon Press.

Habermas, J. (1989). The structural transformation of the public sphere (T. Burger, Trans.). MIT Press. (Original work published 1962)

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). Algorithmic amplification of politics on Twitter. Proceedings of the National Academy of Sciences, 119(1), e2025334119. https://doi.org/10.1073/pnas.2025334119

Jadhav, R.K., E, S., Kurulekar, M., Goel, P., Bhat, U., Upadhyay, M. (2026). Automated Editing Tools for Media Students: A Comparative Study. ShodhKosh: Journal of Visual and Performing Arts, 7(1s), 107–116. doi: 10.29121/shodhkosh.v7.i1s.2026.7075

Klonick, K. (2023). The governance of online speech: Platform constitutionalism and its discontents. Yale Law Journal, 132(6), 1601–1662. https://doi.org/10.2307/yjlf.132.6.1601

Noble, S. U. (2021). Algorithms of oppression: How search engines reinforce racism (Revised ed.). NYU Press.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). McGraw-Hill.

Poell, T., Nieborg, D. B., & Duffy, B. E. (2022). Platforms and cultural production. Polity Press.

Rashmi, C. P., & Jain, L. (2024). Visual Aesthetics and Cinematic Techniques in Indian Mythological Films: An In-Depth Exploration. International Journal of Media and Information Literacy, 9(2), 413-423.

Rashmi, C. P., Jain, M. L., Saroj, N., Bhavsar, R., & KP, Z. (2025). The Impact of Augmented Reality (AR) on Television Advertising: A Consumer Perspective. Advances in Consumer Research, 2, 4073-4085.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2023). Auditing radicalization pathways on YouTube. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 131–142. https://doi.org/10.1145/3593013.3593988

Roth, V., & Pickard, V. (2022). Democracy without journalism? Confronting the misinformation society (Updated ed.). Oxford University Press.

Shin, D., & Biocca, F. (2021). Explicability, causability, and algorithmic transparency: The mediating role of perceived fairness in shaping user trust. Information, Communication & Society, 24(14), 2074–2094. https://doi.org/10.1080/1369118X.2021.1986070

Suzor, N. P. (2021). Lawless: The secret rules that govern our digital lives (Paperback ed.). Cambridge University Press.

Suzor, N. P., West, S. M., & Quodling, A. (2022). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. Social Media + Society, 8(3), 1–13. https://doi.org/10.1177/20563051221123086

Van Dijck, J., Poell, T., & de Waal, M. (2018). The platform society: Public values in a connective world. Oxford University Press.

Zarouali, B., Brosius, A., Helberger, N., & de Vreese, C. H. (2022). Using a 'populist news diet' to explain exposure to and the effects of populist attitudes. New Media & Society, 24(3), 599–618. https://doi.org/10.1177/1461444820946455