

Algorithmic Content Moderation and Transparency in Social Media: Distinguishing Systemic and Procedural Transparency in Regulatory Practice

Tanishq Abdus Subhan

Student Researcher, Loyola International School, Qatar

Jasreet Kaur

Undergraduate Researcher, Bachelor of Science in Analytics, Golden Gate University, USA

Dr. Smrite Goudhaman

Global Professor of Practice, Golden Gate University

Abstract

Recommendation and content-moderation algorithms now govern most of what people encounter on social media, replacing simple chronological feeds with systems that rank, demote, and remove content largely out of public view (Gorwa et al., 2020). This article examines why that opacity has become a policy flashpoint, distinguishing between the curatorial function of ranking and recommendation algorithms and the removal function of content moderation, while noting that both raise related concerns about bias, inconsistency, and the limited ability of affected users to contest decisions (Klonick, 2018). It surveys the leading regulatory responses to this problem, focusing on the European Union's Digital Services Act and its enforcement record, the more fragmented and constitutionally constrained approach taken in the United States, the United Kingdom's (UK) risk-based Online Safety Act, and a contrasting judicial route to platform accountability recently taken by Brazil's Supreme Court. It then turns to the central disagreement in the field: whether mandating disclosure of algorithmic design choices helps or harms the information environment, and whether structured, audited access for vetted researchers offers a workable middle ground between full disclosure and none (Suzor et al., 2019). Drawing on this literature, the article proposes a modest distinction between systemic transparency (disclosure about how a system works in general) and procedural transparency (disclosure to an individual about a specific decision), and argues that current regulation is comparatively well developed on the former and underdeveloped on the latter. The article concludes that transparency is not a single policy lever but a bundle of distinct obligations, that enforcement rather than the mere existence of legislation determines real-world impact, and that procedural transparency toward individual users appears to matter for perceived legitimacy in ways that current regulation has not fully addressed (Jhaver et al., 2019).

Keywords: algorithmic transparency; content moderation; social media regulation; Digital Services Act; platform accountability

1. Introduction

Every click on Instagram, every scroll through X, and every TikTok video is subjected to hundreds of algorithmic decisions about what to show, hide, rank, and never show at all. These systems have replaced the simple reverse-chronological feeds of the early social web, and alongside them sit content-moderation algorithms that flag, demote, or outright remove posts before most users ever see them (Gorwa et al., 2020). Klonick (2018) describes the platforms that operate these systems as “new governors” of online speech: private entities that have developed detailed, rule-based systems for regulating expression at a scale historically reserved for states, but largely outside public view. The rise in efficiency that algorithmic governance enables has come hand in hand with a rise in opacity.

Social media is a primary source of news, community, and debate for many people (St. Aubin & Liedke, 2025), which is why the question of how such systems decide what we see, and who is aware of that, has moved from a niche technical concern to one that now draws policy fights from governments, researchers, journalists, and the platforms themselves.

Research Questions

- **RQ1.** What do the leading regulatory responses to this problem, particularly the EU's Digital Services Act, reveal about which mechanisms actually change platform behavior?
- **RQ2.** How can policymakers reconcile competing demands for disclosure and its limits, and does the distinction between systemic and procedural transparency help resolve that tension?

2. Methodological Framework and Scope

This manuscript is a structured policy and literature review rather than an empirical study or a formal systematic review. It synthesizes scholarship from communication studies, computer science, and legal studies alongside primary regulatory documents and enforcement decisions.

2.1 Search Strategy and Selection Criteria

To ensure reproducibility, sources were retrieved through targeted database searches (including Google Scholar, Scopus, Westlaw, and EUR-Lex) conducted between January and March 2026. The search window covered material published between 2014 and 2026, capturing the evolution from early platform self-governance following major data scandals (2014–2018) through to active statutory enforcement regimes (2022–2026). Key search terms included combinations of algorithmic transparency, algorithmic content moderation, recommender systems ethics, content moderation appeals, procedural justice platforms, researcher data access, and the name of each regulatory instrument discussed in Section 4.

The reference list comprises 28 sources across four categories: 14 peer-reviewed academic articles or conference papers spanning communication studies, human-computer interaction, and

philosophy of technology; 4 primary legislative or regulatory texts (the EU DSA, the UK OSA, California SB 771, and US S. 2024); 7 official judicial and regulatory records documenting specific enforcement actions or legislative history (European Commission non-compliance and commitment decisions, Ofcom confirmation decisions and investigation notices, the Brazilian Supreme Court judgment, the General Court of the European Union’s case notice for X’s appeal, and the official legislative record of SB 771’s veto); and 3 legal-analysis or policy-commentary sources used to document contested or fast-moving developments for which no peer-reviewed literature yet exists (Linklaters, the Global Network Initiative, and the Association for Competitive Technology).

2.2 Inclusion and Exclusion Protocol

A source was included if it addressed algorithmic transparency or accountability in content moderation or recommendation systems directly, or if it was the primary legal text, court judgment, regulatory decision, or first-hand procedural record for a jurisdiction discussed below. Table 1 sets out the full inclusion and exclusion protocol applied during source selection.

Table 1

Source Inclusion and Exclusion Protocol for Literature and Regulatory Review

Criterion	Inclusion Criteria	Exclusion Criteria
Domain Scope	Direct focus on social media algorithmic curation, ranking, or automated moderation	Algorithmic governance in non-media contexts (e.g., credit scoring, criminal sentencing, employment)
Document Type	Peer-reviewed journal articles or conference papers, law review articles, enacted statutes and bills, formal regulatory decisions, court judgments, and official legislative records	Unverified news blogs, non-peer-reviewed commentary, or secondary journalistic summaries without a traceable primary source
Temporal Range	Published between 2014 and March 2026	Material published before 2014

Criterion	Inclusion Criteria	Exclusion Criteria
Language	English language, or official translations of court judgments (e.g., the Brazilian STF)	Non-translated foreign-language texts lacking verified translations

Note: This protocol structures the source selection for conceptual synthesis; it does not constitute a formal PRISMA-style systematic review.

This is a purposive rather than systematic sample. It does not claim to capture the full literature in any of the five subfields it draws on, and a systematic review of any one of them—platform governance, recommender-system ethics, content-moderation appeals, researcher data access, or user trust in automated decisions—would merit a dedicated article of its own.

2.3 Conceptual Scope

The unit of analysis is the platform's algorithmic governance system as a whole, encompassing both its recommendation and ranking functions (curation) and its content-moderation functions (removal), rather than any single model or piece of legislation in isolation. Where the article distinguishes between systemic transparency, disclosure about how a system works in general, and procedural transparency, disclosure to an individual about a specific decision affecting them, this distinction is offered as an organizing lens for the literature and regulatory record reviewed below, not as an empirically validated model. Table 2 below sets out both terms side by side before they are used throughout the rest of the article.

Table 2

Conceptual Comparison of Systemic and Procedural Transparency

Dimension	Systemic Transparency	Procedural Transparency
Definition	Disclosure about how a system works in general: its rules, design, or aggregate behavior	Disclosure to an individual about a specific decision that affected them
Example	An advertising repository; a published moderation policy; a platform-wide algorithmic risk assessment	A removal notice explaining which specific rule a given post violated

Dimension	Systemic Transparency	Procedural Transparency
Intended Audience	Regulators, researchers, and the public	The individual user affected by the decision
Purpose	Enable oversight of the system as a whole; help detect patterns of bias or harm	Enable the affected person to understand, and potentially contest, a specific outcome
Possible Regulatory Mechanism	Mandatory disclosure and reporting duties, audits, researcher data-access rights (e.g., DSA Art. 40)	Statement-of-reasons requirements, individual notice duties, and appeal rights (e.g., DSA Arts. 17, 20–21)

Note: Framework adapted from conceptual distinctions synthesized across platform governance literature (e.g., Ananny & Crawford, 2018; Suzor et al., 2019).

3. Literature Review

3.1 Algorithmic Moderation and Recommendation as Related but Distinct Functions

Content moderation on major platforms is no longer purely a human endeavor. Gorwa et al. (2020) describe how machine learning systems scan uploads in real time and compare text, images, and video against policy rules on hate speech, graphic violence, misinformation, spam, and more. Human review remains part of the process, particularly for edge cases and appeals, but Gorwa et al. (2020) note that automated tools have become the first line of defense against prohibited content, and increasingly the last.

Recommendation and ranking algorithms are related to content moderation but distinct from it: they decide what gets shown in a user's feed at all. Milano et al. (2020), in a systematic review of the ethics of recommender systems, identify six areas of ethical concern, including manipulation, opacity, and the unequal treatment of different user groups, and argue that most existing scholarship evaluates these systems only from the recipient's point of view, overlooking the interests of other stakeholders such as content creators and advertisers. This matters for transparency specifically: a recommendation system can be technically compliant with disclosure rules aimed at users while remaining opaque to the other parties its rankings affect. The two functions, curation and removal, are structurally different, but they raise similar concerns about how they shape perception and behavior. Klonick (2018) situates both within a broader account of platforms as private governance systems that combine free-expression norms, corporate responsibility, and commercial incentives, developing rules and enforcement mechanisms that

resemble common-law regulation without the accountability structures that typically accompany public governance.

Diakopoulos (2015) offers a complementary account from outside platform governance scholarship: algorithmic accountability reporting, in which journalists and outside investigators reverse-engineer a system's inputs and outputs to infer how it behaves, has functioned as a partial substitute for formal disclosure when platforms decline to explain their systems directly. This tradition is a reminder that transparency obligations are not the only route to accountability, though Diakopoulos (2015) also notes that reverse engineering is resource-intensive and produces, at best, an approximate picture of a system's actual logic.

3.2 The Transparency Ideal and Its Limits

Public concern about the opacity of social media algorithms has not arisen in a vacuum. Data misuse scandals, accusations of suppression or elevation of particular viewpoints, and documented links between recommendation systems and the spread of misinformation and extremist content have driven lawmakers and researchers to take an interest in algorithmic transparency, if not always in practice then at least in principle. Ananny and Crawford (2018) caution that transparency is not a straightforward remedy: seeing a system, in the sense of receiving disclosure about it, does not necessarily mean understanding how power operates through it. Disclosure can leave the underlying institutional relationship unchanged, particularly when the recipient lacks the expertise, time, or standing to act on the information.

Suzor et al. (2019) push this critique toward specificity, arguing that transparency is not one thing but refers to several distinct questions: the selection, ranking, and recommendation logic of a system; the moderation and policy practices to which content is subjected; the data these systems use as input; and the guidance under which they operate. Based on a survey of individuals who had experienced content moderation decisions, they argue that meaningful transparency should be understood as part of a communicative process of accountability to the people affected by a decision, not merely the publication of aggregate statistics. Regulators, researchers, journalists, and the public each need different kinds of access to investigate these questions, which is part of what makes “just be transparent” such a neat but ultimately underspecified policy directive.

The procedural side of this distinction has a legal history that predates the DSA and helps explain why it is not, on its own, a new idea. Articles 13 through 15 and 22 of the EU's General Data Protection Regulation (GDPR) grant individuals subject to a solely automated decision with legal or similarly significant effects a right to “meaningful information about the logic involved” in that decision. Selbst and Powles (2017) argue, against an influential contrary reading, that this language does amount to a right to explanation, and that it should be interpreted functionally and flexibly enough to let a person actually exercise their other rights under the GDPR, such as contesting a decision or seeking human review. The contrary reading they are responding to, Wachter, Mittelstadt, and Floridi (2017), argues that the GDPR's operative articles do not, as a matter of legal text, establish a binding right to an individual explanation at all, and that any such right exists only weakly, in a non-binding recital. That the two most cited legal analyses of the same provisions reach opposite conclusions is itself informative: even in the jurisdiction with the

most developed data-protection law in the world, whether individuals are legally entitled to an explanation of an automated decision affecting them remains unsettled eight years after the GDPR took effect. That debate, over what counts as an adequate individual-level explanation of an automated decision, is functionally the same question this article calls procedural transparency, applied a decade earlier to automated decision-making in general rather than to content moderation specifically. Content moderation and recommendation decisions do not always meet the GDPR's threshold of a solely automated decision with legal or similarly significant effects, which is one reason the DSA's own statement-of-reasons requirement in Article 17 exists as a separate, more specific obligation. But the underlying doctrinal question, what does an individual need to be told for a decision about them to be genuinely contestable, is the same one, and situates procedural transparency as an application of an existing legal debate rather than a freestanding concept invented for this article.

3.3 Inconsistency and Arbitrariness in Automated Enforcement

A further complication is that algorithmic moderation decisions are not always consistent, even for materially similar content. Gomez et al. (2024) show that predictive multiplicity, the phenomenon in which multiple machine learning models achieve similar average accuracy but disagree on individual cases, is common in content-moderation classifiers for toxic text. Because platforms rarely disclose which of several equally accurate models made a given decision, or why, similar posts can receive different outcomes for reasons unrelated to their content. Gomez et al. (2024) connect this arbitrariness to concerns about freedom of expression, non-discrimination, and procedural justice, arguing that it is not simply a source of error to be minimized but a structural feature of how these systems are built and deployed.

3.4 Contestation, Appeals, and Researcher Access

A separate strand of research asks what happens after a moderation decision is made, rather than how the decision itself was reached. Vaccaro et al. (2020), studying how Facebook users experienced the process of appealing a content removal, found that formal appeal mechanisms did not, on their own, improve users' perceptions of the fairness, accountability, or trustworthiness of the underlying decision. Users who filed appeals often contested not just the individual outcome but more fundamental issues, including the goal of moderating certain content at all, the use of automation in the first place, and the perceived inconsistency of the system as a whole. This suggests that appeal mechanisms, though frequently proposed as a remedy for opaque moderation, do not by themselves resolve the transparency problem unless paired with a genuine explanation of the original decision.

A related and largely separate line of work concerns transparency toward researchers rather than toward individual users. Ohme and Seiling (2026) examine how Article 40 of the DSA, which grants vetted researchers a legal right to request data from very large platforms, has functioned in its first years of operation. They find that the provision represents a genuine structural change from the prior status quo of purely voluntary data sharing, but that early implementation has been uneven, with unclear technical standards and practical limits on what platforms will share even under a formal legal request. Researcher access is therefore a third transparency channel, distinct

from both the systemic disclosure aimed at the public and the procedural explanation owed to an individual user, and one that the regulatory record examined in Section 4 shows is still being tested in practice.

3.5 Procedural Transparency and User Trust Across Platforms

Whether individual users are told why their content was moderated shapes how they respond to that decision. Jhaver et al. (2019), surveying Reddit users shortly after their posts were removed, found that over a third did not understand why their post had been removed, and that users who received an explanation or had read a community's rules in advance were substantially more likely to view the removal as fair and to continue participating in the community. Because this finding comes from a single platform built around community-specific rules, it is worth asking whether it generalizes. Myers West (2018), surveying 519 users across multiple mainstream platforms who had experienced content moderation, found a similar pattern: users routinely developed their own folk theories about why content had been removed in the absence of clear explanation, and these theories shaped a largely adversarial relationship with the platform rather than a cooperative one. Myers West (2018) argues for an educational rather than purely punitive model of moderation, in which platforms explain decisions in a way that helps users understand the rule in question, not merely that a rule was broken. Read together, the Reddit-specific and cross-platform findings support the distinction, introduced above, between systemic transparency, disclosure about how a system works in general, and procedural transparency, disclosure to an individual about a specific decision: the latter appears to matter directly for perceived legitimacy across platform contexts, independent of how much a platform discloses about its policies overall.

3.6 Research Gap and Contribution

Individually, the literature reviewed above already establishes several of this article's building blocks: that algorithmic transparency is not one thing but several distinct questions (Suzor et al., 2019), that transparency as a policy ideal has real limits (Ananny & Crawford, 2018), that moderation decisions can be inconsistent in ways that policy disclosure alone does not fix (Gomez et al., 2024), that formal appeal mechanisms do not by themselves restore user trust (Vaccaro et al., 2020), and that individual explanation of a specific decision changes how users respond to it, a pattern observed on Reddit and confirmed across a larger, multi-platform sample (Jhaver et al., 2019; Myers West, 2018).

What this literature does not do, and what this article adds, is connect these separate findings to the regulatory record by asking a narrower and more concrete question: when governments actually legislate for algorithmic transparency, which of these several things do they legislate for? The systemic and procedural transparency distinction used throughout this article is not, on its own, a new theoretical concept; the underlying idea, that transparency can concern a system in general or a specific decision affecting one person, is already present, in different terms, across the literature reviewed in this section. The contribution here is better described as a synthesis and an organizing lens rather than a new theory: naming that distinction explicitly and using it to read the EU, U.S., UK, and Brazilian regulatory records side by side, as in Table 3

below, surfaces a pattern that is not stated directly in any single source reviewed here. Binding law across these jurisdictions is comparatively well developed on systemic disclosure, and comparatively underdeveloped on procedural, individual-level explanation; the user-trust literature reviewed in Section 3.5 suggests that this gap matters for perceived legitimacy, though that literature speaks to user trust rather than to a direct comparison against the effects of systemic disclosure. This article's contribution is accordingly a synthesis rather than a new empirical finding: it is the pattern surfaced by reading the regulatory record through this lens, not the systemic/procedural terminology itself, that the rest of the article develops.

4. Regulatory Approaches Across Jurisdictions

4.1 The European Union's Digital Services Act

The farthest-reaching effort to legislate algorithmic transparency comes in the form of the EU's Digital Services Act (DSA), Regulation (EU) 2022/2065, which took effect in 2022 (European Parliament and Council of the European Union, 2022). The DSA imposes special obligations on “Very Large Online Platforms” and “Very Large Online Search Engines,” defined as those reaching more than roughly 10% of the EU's population, to identify, assess, and mitigate systemic risks connected to their algorithmic systems, including risks to civic discourse, elections, and the mental health and safety of minors.

Enforcement has had teeth, though it is not yet final. On 5 December 2025, the European Commission fined X €120 million in the DSA's first non-compliance decision, finding that the platform's verification badge was deceptively presented, that its advertising repository fell short of disclosure requirements, and that it had made it unduly difficult for researchers to access public data (European Commission, 2025a). X filed an appeal against the fine with the General Court of the European Union on 16 February 2026, contesting both the underlying findings and the Commission's procedure; as of this writing, that appeal is pending and has been reported as the first judicial challenge to a DSA fine (General Court of the European Union, 2026). On the same day, the Commission announced that TikTok had submitted binding commitments, enforceable for five years, to bring its advertising repository into compliance after the Commission's preliminary finding in May 2025 that the repository did not meet DSA transparency requirements; that case was resolved through commitments rather than a fine (European Commission, 2025b). Both cases were procedurally systemic rather than individual: they concerned platform-wide disclosure obligations rather than the treatment of any specific user or post.

4.2 The United States

Efforts in the U.S. have been inconsistent and slower to produce binding law, constrained by Section 230 protections and First Amendment concerns that limit the government's ability to directly regulate content or algorithmic design. The bipartisan Filter Bubble Transparency Act, first introduced in the Senate in 2019 and reintroduced in subsequent Congresses, would require large platforms to let users opt into a feed that is not curated by a personal-data-driven algorithm (U.S. Congress, Senate, 2021). The bill illustrates the broader difficulty of writing transparency rules that do not also degrade features users value, such as location-based recommendations, and it has not been enacted.

At the state level, California's SB 771 would have imposed civil liability on social media platforms whose recommendation algorithms knowingly amplified content that facilitated discrimination or violence against protected groups (California State Senate, 2025). The bill passed both chambers of the state legislature in 2025 but was vetoed by the governor, and on 2 March 2026 the Senate voted to sustain the veto rather than override it (California State Legislature, 2026), illustrating how quickly a state-level transparency or liability effort can stall even after clearing the legislature. Together, these efforts represent a patchwork regulatory environment in the absence of comprehensive federal legislation.

4.3 The UK and Elsewhere

The UK's Online Safety Act 2023 takes a different approach, focusing on platform accountability for illegal and harmful content, with algorithmic systems folded into a broader risk-assessment framework rather than made the direct subject of standalone transparency duties (UK Parliament, 2023). Platforms must maintain accessible complaints mechanisms and explain how moderation decisions are made, and Ofcom can fine noncompliant services up to £18 million or 10% of qualifying worldwide revenue, whichever is greater. Enforcement began in earnest in mid-2025, with Ofcom opening its first investigations of individual services, including investigations into Kick Online Entertainment S.A. — the corporate operator of the pornography site Motherless.com — over alleged illegal-content and age-assurance compliance failures (Ofcom, 2025c).

The 4chan case has gone furthest and illustrates a distinctly UK enforcement problem largely absent from the DSA record. After 4chan failed to respond to statutory information requests, Ofcom issued a Confirmation Decision and, by April 2026, had imposed fixed penalties of £520,000 plus daily accumulating fines for continued non-compliance (Ofcom, 2026); 4chan has refused to pay and continues to dispute Ofcom's jurisdiction. In August 2025, 4chan and the discussion board Kiwi Farms had already filed suit in a US federal court seeking to block Ofcom from enforcing the Act against services with no UK physical presence, underscoring the jurisdictional limits facing a national, rather than EU-wide, regulator when a platform simply declines to comply (Linklaters, 2025). Content moderation legislation has been applied inconsistently around the world, with jurisdictions balancing free expression, platform liability, and user safety in different ways. As a result, global platforms increasingly must apply differing moderation and disclosure regimes across jurisdictions and regions.

4.4 Brazil: A Contrasting Judicial Model

The EU, US, and UK cases above share a common logic even where they disagree: transparency, in whatever form, is meant to be legislated or regulated into existence. Brazil's recent experience shows that a jurisdiction can arrive at serious platform accountability through an entirely different route, and that route can bypass transparency obligations altogether. Since 2014, Brazil's Marco Civil da Internet had shielded platforms from liability for user-generated content unless they ignored a specific court order to remove it, a safe-harbor model similar in spirit to Section 230 in the United States. In June 2025, Brazil's Federal Supreme Court (STF) ruled, by an 8–3 majority, that this safe harbor was partially unconstitutional, holding that platforms can now be held civilly liable for certain third-party content even without a prior court order, particularly for serious crimes and, under a fault-based standard, where a platform's systemic failures allowed harm to spread (Brazilian Federal Supreme Court, 2025).

What makes the Brazilian case a useful contrast is what it does not contain. The ruling was reached judicially rather than legislatively, imposes real liability with no cap comparable to the DSA's or Ofcom's fine structures, and yet includes no disclosure duty, no individual statement-of-reasons requirement, no researcher-access provision, and no appeal mechanism of the kind found in the DSA or the OSA (Global Network Initiative, 2025). It substitutes a general duty of diligence—platforms must act with the reasonable care a court will later assess after the fact—for the ex ante transparency obligations that structure the EU and, on paper, the UK approach. This suggests that transparency and accountability, though often discussed together, are not the same lever: a jurisdiction can impose severe consequences for platform failure while telling the public and affected users comparatively little about how moderation and ranking decisions are made in the first place. Whether Brazil's model produces better outcomes for users than the EU's disclosure-heavy approach is an open empirical question the case law is too recent to answer, but the comparison itself demonstrates that the EU's specific bundling of disclosure and liability is a policy choice, not the only available one.

4.5 Comparative Matrix Across Four Jurisdictions

Table 3 compares the four jurisdictions directly against six criteria drawn from the literature and regulatory record reviewed above: systemic disclosure, individual notice and explanation, appeal and redress, researcher access, enforcement powers, and major limitations.

Table 3

Comparative Analysis of Platform Regulatory Frameworks Across Four Jurisdictions

Criterion	European Union (DSA)	United States	United Kingdom (OSA)	Brazil (STF Jurisprudence)
Systemic disclosure	Mandatory for Very Large Online Platforms and Search Engines: risk assessments, advertising repositories, and algorithm disclosure obligations (European Parliament and Council of the European Union, 2022)	No binding federal requirement; the Filter Bubble Transparency Act proposed opt-out disclosure but was never enacted (U.S. Congress, Senate, 2021)	Illegal- and harmful-content risk assessments required, but not algorithmic ranking disclosure as a standalone duty (UK Parliament, 2023)	No statutory or judicial systemic disclosure duties mandated (Global Network Initiative, 2025)
Individual notice / explanation	Statement of reasons required for moderation decisions (DSA Art. 17)	No federal requirement; disclosure to individual users is voluntary and platform-specific	Platforms must give clear information about moderation decisions, but with no codified	No explicit requirement for individual statement-of-reasons or moderation explanation

Criterion	European Union (DSA)	United States	United Kingdom (OSA)	Brazil (STF Jurisprudence)
			statement-of-reasons right comparable to the DSA	
Appeal / redress	Internal complaint-handling (Art. 20) plus certified out-of-court dispute settlement bodies, e.g., Appeals Centre Europe (Art. 21)	No standardized or legally mandated mechanism; entirely platform-specific	Platforms must provide accessible internal complaints mechanisms, but no independent, DSA-style dispute-settlement body exists	Redress occurs primarily through direct judicial channels and civil liability claims
Researcher access	Legal right for vetted researchers under Art. 40, though implementation has been uneven in practice (Ohme & Seiling, 2026)	No legal right; access is voluntary and contractual, and can be revoked by the platform	No equivalent statutory provision; Ofcom has investigatory powers but no researcher-data-access mandate	No statutory or judicial framework providing mandated researcher data access
Enforcement powers	Fines up to 6% of global turnover; first fine issued against X in December 2025, now under appeal	SB 771 would have created civil liability in California but was vetoed; no comprehensive federal	Fines up to £18 million or 10% of qualifying worldwide revenue; first individual-service	Judicial civil liability uncapped by statute; fault-based liability imposed for systemic failures and serious

Criterion	European Union (DSA)	United States	United Kingdom (OSA)	Brazil (STF Jurisprudence)
	at the EU General Court; TikTok case resolved via binding commitments (European Commission, 2025a, 2025b)	enforcement regime exists (California State Senate, 2025)	enforcement (4chan) has resulted in fixed penalties of £520,000 plus daily accumulating fines, still unpaid amid open defiance and jurisdictional litigation in US courts (Ofcom, 2026; Linklaters, 2025)	illegal content (Brazilian Federal Supreme Court, 2025)
Major limitations	Disclosure concentrated at the systemic level; out-of-court dispute bodies are new and untested at scale; researcher access implementation inconsistent	Fragmented and largely unenacted; Section 230 and First Amendment constraints limit direct regulation of content or algorithmic design	Focused on illegal and harmful content rather than algorithmic transparency per se; no researcher-access mandate; enforcement against non-UK platforms is jurisdictionally contested	Ex-post judicial liability model lacks proactive ex-ante transparency, statement-of-reasons, or structured researcher oversight mechanisms

Note: Data compiled from legislative texts, regulatory decisions, and court records as cited in Section 4.

5. The Core Debate: How Much Transparency Is Too Much?

Not everyone agrees that increased disclosure of algorithmic design choices is a net benefit. Critics have warned that mandating specific design choices rather than outcomes could grant governments too much control over the information environment (Klonick, 2018). There are also practical limits to the technical feasibility of full disclosure, particularly for recommendation and ranking systems: Ananny and Crawford (2018) note that disclosure alone can overwhelm rather than inform when the recipient lacks the capacity to interpret it. Others argue that full disclosure could be exploited by bad actors seeking to manipulate the moderation system (Veale, Van Kleek, & Binns, 2018), and that overly broad mandates could be unworkable for smaller platforms that lack the capacity of a Meta or a Google (Association for Competitive Technology, 2021).

Proponents of disclosure point to platforms' limited record of internal self-correction as a reason to adopt stronger third-party oversight. Suzor et al. (2019) suggest that structured, audited access is one workable alternative to full disclosure, a view echoed elsewhere in this literature: giving vetted, independent researchers and regulators the means to evaluate systems for bias and harm without exposing implementation details that could create new vulnerabilities.

A further question of practicality lies within moderation practices themselves. Gomez et al. (2024) show that moderation decisions can be inconsistent even for similar content, a problem that systemic disclosure of policies does little to address because the inconsistency arises from model-level arbitrariness rather than from unclear rules. Jhaver et al. (2019) complement this finding from the user's side: explanations of individual decisions changed how people reacted to a removal, indicating that procedural transparency, being told why a specific post was removed, appears to matter directly for perceived legitimacy, though this finding speaks to user trust rather than to a like-for-like comparison against the effects of systemic disclosure.

6. Where This Leaves Platforms, Users, and Regulators

6.1 Answering the Research Questions

RQ1 asked what the leading regulatory responses, particularly the EU's DSA, reveal about which mechanisms actually change platform behavior. The regulatory record in Section 4 and Table 3 offers a more mixed answer than a simple “fines work” conclusion, and the two DSA cases usefully illustrate different points on that spectrum. TikTok altered its advertising repository and accepted binding, five-year commitments after the Commission issued preliminary findings of non-compliance, before any fine was imposed, suggesting that the credible threat of enforcement, not a fine itself, can be sufficient to produce a negotiated change in practice (European Commission, 2025b). X, by contrast, was fined €120 million and has not visibly altered the practices the Commission cited; instead, it filed an appeal with the EU General Court on 16 February 2026 disputing the findings and the Commission's procedure (General Court of the European Union, 2026), so whether the fine itself changes X's behavior remains genuinely

unresolved at the time of writing (European Commission, 2025a). Read together, the two cases suggest that a credible threat of enforcement can shift behavior even without a final penalty, while an imposed and contested fine does not guarantee compliance on its own. By contrast, the U.S. Filter Bubble Transparency Act and California's SB 771 show that transparency or liability rules without a clear enacted legal basis produce no behavioral change at all, however well designed (U.S. Congress, Senate, 2021; California State Senate, 2025). The UK's Online Safety Act suggests a further, jurisdictional condition: even a binding law with escalating real penalties changes behavior only for platforms within the regulator's practical reach, as 4chan's continued non-payment and litigation against Ofcom illustrates (Linklaters, 2025).

RQ2 asked how policymakers can reconcile competing demands for disclosure and its limits, and whether the systemic and procedural transparency distinction developed in Section 3.6 helps resolve that tension. The core debate in Section 5 shows that the strongest objections to disclosure, that it could be exploited by bad actors or overwhelm smaller platforms, apply mainly to systemic disclosure of design choices, model weights, or ranking logic. Those objections apply far less to procedural transparency: telling one user which specific rule their post violated does not hand a blueprint to bad actors in the way that publishing full ranking logic might, and the evidence reviewed in Section 3.5 suggests it delivers a comparatively large gain in perceived legitimacy for a comparatively small disclosure risk. The distinction therefore offers a partial answer to RQ2: rather than treating “more transparency or less” as a single dial, policymakers can expand procedural transparency, which carries fewer of the risks critics raise, even where they remain cautious about further systemic disclosure.

6.2 Synthesis, Effectiveness, and Practical Implications

Several conclusions follow when the literature and regulatory record in Sections 3 and 4 are read together. Transparency is not a single policy lever but several: disclosure of ranking criteria, documentation of moderation policy, researcher access to data, and explanation of individual removals all have their place, but require different mechanisms and levels of oversight. Table 3 makes this concrete: the EU is comparatively strong on systemic disclosure, appeal and redress, and researcher access, while the U.S. lacks binding requirements on all six criteria, the UK is strong on enforcement powers but has no researcher-access mandate and a weaker individual-explanation right than the DSA, and Brazil focuses on civil liability without ex-ante disclosure duties. As the TikTok, X, and 4chan cases discussed under RQ1 above illustrate, legal and regulatory attention so far has concentrated on systemic disclosure rather than procedural, individual-level explanation, even though the evidence in Section 3.5 suggests the latter matters directly for perceived fairness.

It is enforcement, rather than the mere existence of legislation, that determines a law's real-world impact, and the honest answer on that score is mixed for every jurisdiction examined. The EU's disclosure-heavy approach has produced the field's clearest behavioral result, TikTok's negotiated commitments, but its flagship enforcement action against X is stalled in litigation that could take years to resolve, during which the disclosure failures the fine was meant to correct remain uncorrected: a disclosure mandate backed by a penalty is not self-executing, and a determined platform can convert it into a multi-year legal question rather than an operational change. The UK's experience is starker still: Ofcom has real statutory fining power, has used it

against 4chan, and has still collected almost nothing, because a transparency regime with strong formal powers but no practical reach against a noncompliant, foreign-based platform is, for that platform, functionally optional. The U.S. record is more straightforward but no more encouraging, since without enacted federal law there is no binding obligation to test. Brazil's contrasting model (Section 4.4) complicates the picture further: the STF's ruling achieves a real, judicially backed consequence for platform failure without any transparency apparatus at all, but a liability rule enforced after harm has occurred is a different instrument from a disclosure rule meant to let regulators, researchers, and users see a problem before it escalates. Taken together, no single mechanism—disclosure duty, statement of reasons, researcher access, or after-the-fact liability—reliably produces both real accountability and real transparency at once; each model examined here achieves at most one or two of the six criteria in Table 3 in practice rather than merely on paper.

Two practical implications follow. For platforms, the systemic/procedural distinction points to an underexploited option: the cheaper, less legally contested half of the transparency agenda, aggregate disclosure, has drawn more regulatory conflict than the comparatively less contested half, individual explanation, despite evidence that the latter does more to shape user trust (Section 3.5); investing in clearer, more specific individual notices, of the kind Myers West (2018) and Jhaver et al. (2019) both associate with higher perceived fairness, is less likely to trigger the kind of prolonged legal contest that systemic disclosure mandates like the DSA's have produced, for a comparatively large trust benefit. For regulators, the four-jurisdiction comparison suggests that a disclosure regime is only as strong as its enforcement reach: statutory fining power aimed at platforms with no domestic presence needs a credible collection or access-restriction mechanism behind it, and regulators enforcing a first-of-its-kind fine, as the EU is discovering with X, should expect and plan for extended legal contest as a normal part of the process rather than treating the initial decision as the end of it. Brazil's example also suggests that transparency and liability need not be bundled together as they are in the DSA; a regulator prioritizing accountability quickly, with less institutional capacity to build a full disclosure apparatus, has a demonstrated alternative in strengthening after-the-fact liability rather than waiting to build an EU-style systemic transparency regime from scratch.

7. Limitations

This article has three main limitations. First, it is conceptual rather than empirical: the systemic/procedural transparency distinction is offered as an organizing lens for existing findings, not as a model that has itself been tested. Second, the literature and regulatory record reviewed are integrative rather than systematic, and are concentrated on the EU, US, UK, and Brazil; substantial regulatory activity in other jurisdictions, including much of Asia, Africa, and the rest of Latin America, is outside the scope of this article. Third, the regulatory record discussed is recent and, in the case of the DSA enforcement actions and SB 771's veto, still unfolding at the time of writing; the longer-term effects of these actions on platform behavior are not yet observable.

8. Future Research

Future work could empirically test whether platforms that provide clearer procedural explanations, of the kind studied by Jhaver et al. (2019) in the Reddit context, see measurably different user trust or engagement outcomes on larger commercial platforms. Comparative research could examine whether DSA enforcement actions produce durable changes in platform disclosure practices beyond the specific violations cited, or whether compliance remains narrowly targeted to the terms of each decision. Finally, given that most binding transparency law reviewed here addresses systemic rather than procedural disclosure, research and policy design attention to individual-level explanation requirements, and to their effect on perceived legitimacy, appears to be a comparatively underexplored area.

9. Conclusion

Algorithmic content moderation and recommendation systems have scaled public discourse beyond anything previously possible under human curation. That same scale is the reason why questions of transparency have become such contentious policy issues: the systems that make the modern social media environment possible also make its decision-making largely invisible to the people it affects. The regulatory experiments now underway, from the EU's audited-access model to the patchwork approach in the U.S. and the UK's risk-based framework, are ultimately a test of how democratic societies can hold automated decision-making accountable without either paralyzing platforms or granting governments undue control over what people can see online. The distinction between systemic and procedural transparency developed here suggests that the next stage of that experiment may need to focus less on what platforms disclose about their systems in general, where the DSA has already made real progress, and more on what individual users are told about the decisions that affect them directly.

Author Contributions and AI Statement

Correspondence: Correspondence concerning this article should be addressed to Tanishq Abdus Subhan, Loyola International School, Qatar.

Author Contributions: T.A.S. conceptualized the review, synthesized the regulatory and legal record, and authored the primary draft. J.K. performed literature mapping, methodology structuring, and analytical review. S.G. supervised the project, provided theoretical framing, and edited the final manuscript.

AI Use Statement: During manuscript preparation, the authors used an AI assistant (Claude, Anthropic) to conduct literature searches, verify recent regulatory and legal developments, draft and revise text, and format citations and references. The authors reviewed all AI-assisted output, verified sources against original material where possible, and take full responsibility for the content, arguments, and conclusions of this article.

References

- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
- Association for Competitive Technology. (2021, November 10). *How to make the Digital Services Act more SME-friendly*. <https://actonline.org/2021/11/10/how-to-make-the-digital-services-act-more-sme-friendly/>
- Brazilian Federal Supreme Court. (2025). *Recurso Extraordinário 1.037.396 and Recurso Extraordinário 1.057.258 (Temas 987 and 533)* [Judgment of June 26, 2025].
- California State Legislature. (2026, March 2). *Senate Bill No. 771 (2025–2026 Regular Session): Bill status, official record* [includes Governor's veto message of October 13, 2025, and Senate vote to sustain the veto]. https://leginfo.legislature.ca.gov/faces/billStatusClient.xhtml?bill_id=202520260SB771
- California State Senate. (2025). *Senate Bill No. 771, Personal rights: liability: social media platforms (2025–2026 Reg. Sess.)*. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202520260SB771
- Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>
- European Commission. (2025a, December 5). *Commission fines X €120 million under the Digital Services Act* [Press release]. <https://digital-strategy.ec.europa.eu/en/news/commission-fines-x-eu120-million-under-digital-services-act>
- European Commission. (2025b, December 5). *Commission accepts TikTok's commitments on advertising transparency under the Digital Services Act* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2940
- European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act). *Official Journal of the European Union*, L 277. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- General Court of the European Union. (2026, May 4). Action brought on 16 February 2026 — X.AI Holdings v Commission (Case T-120/26) [Notice]. *Official Journal of the European Union*, C/2026/2391. <https://eur-lex.europa.eu/eli/C/2026/2391/oj/eng>
- Global Network Initiative. (2025, November 3). *From shield to scrutiny: Brazil's Supreme Court redefines platform liability*. <https://globalnetworkinitiative.org/from-shield-to-scrutiny-brazils-supreme-court-redefines-platform-liability/>
- Gomez, J. F., Machado, C. V., Paes, L. M., & Calmon, F. P. (2024). Algorithmic arbitrariness in content moderation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2234–2253). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3659036>
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 1–15. <https://doi.org/10.1177/2053951719897945>
- Jhaver, S., Appling, D. S., Gilbert, E., & Bruckman, A. (2019). “Did you suspect the post would be removed?”: Understanding user reactions to content removals on Reddit.

Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), Article 192.
<https://doi.org/10.1145/3359294>

- Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6), 1598–1670.
- Linklaters. (2025, September). *The Online Safety Act 2023 – the landscape two years on*. <https://www.linklaters.com/en/insights/blogs/digilinks/2025/september/uk-the-online-safety-act-2023-the-landscape-two-years-on>
- Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & Society*, 35(4), 957–967. <https://doi.org/10.1007/s00146-020-00950-y>
- Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383. <https://doi.org/10.1177/1461444818773059>
- Ofcom. (2025c, May 14). *Enforcing the Online Safety Act: Ofcom investigates Kick Online Entertainment S.A.* <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/enforcing-the-online-safety-act-ofcom-investigates-kick-online-entertainment-s.a>
- Ofcom. (2026, April 21). *Confirmation Decision: Investigation into 4chan Community Support LLC's compliance with its duties to protect its users from illegal content* [PDF]. <https://www.ofcom.org.uk/siteassets/resources/documents/enforcement/2026/4chan-confirmation-decision.pdf>
- Ohme, J., & Seiling, L. K. (2026). Finally, access: How Article 40 DSA changes platform research in practice. *Political Communication*. <https://doi.org/10.1080/10584609.2026.2664157>
- Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233–242. <https://doi.org/10.1093/idpl/ix022>
- St. Aubin, C., & Liedke, J. (2025, September 25). *Social media and news fact sheet*. Pew Research Center. <https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/>
- Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13, 1526–1543.
- UK Parliament. (2023). *Online Safety Act 2023*, c. 50. <https://www.legislation.gov.uk/ukpga/2023/50/contents>
- U.S. Congress, Senate. (2021). *Filter Bubble Transparency Act*, S. 2024, 117th Cong. <https://www.congress.gov/bill/117th-congress/senate-bill/2024>
- Vaccaro, K., Sandvig, C., & Karahalios, K. (2020). “At the end of the day Facebook does what it wants”: How users experience contesting algorithmic content moderation. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2), Article 167. <https://doi.org/10.1145/3415238>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3174014>
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ix005>

