

 ISSN (O): 2320-5407 ISSN (P): 3107-4928	Journal Homepage: -www.journalijar.com INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR) Article DOI:10.21474/IJAR01/23801 DOI URL: http://dx.doi.org/10.21474/IJAR01/23801	 INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR) Volume 14(07) July-2026 Published Homepage: http://www.journalijar.com Journal ISSN (O) 2320-5407
--	---	---

RESEARCH ARTICLE

CONTENT MODERATION AND FREE SPEECH: LEGAL LIMITS OF AUTOMATED TAKEDOWN SYSTEMS ON SOCIAL PLATFORMS

Kunsang Wangmo

Manuscript Info

Manuscript History

Received: 08 May 2026

Final Accepted: 10 June 2026

Published: July 2026

Key words:-

Automated content moderation; Free speech; Platform governance; Algorithmic arbitrariness; Due process; Human rights law; Copyright takedowns; Online speech regulation.

Abstract

This paper examines automated content moderation as a system of private speech governance and evaluates its compatibility with foundational principles of freedom of expression. Moving beyond technical accounts of algorithmic error, the study conceptualizes platforms as private governors exercising regulatory power through automated, risk-driven enforcement mechanisms. Drawing on legal theory, empirical scholarship, and human rights law, the paper demonstrates how automation restructures moderation into a system marked by procedural failure, substantive arbitrariness, and weak accountability. It argues that opaque notice regimes, ineffective appeals, and biased outcomes are not accidental flaws but structural consequences of scale-oriented automation. Focusing on copyright-based takedown systems as a prototype, the analysis traces the expansion of automation from intellectual property enforcement to general speech control. The paper further evaluates human rights and EU regulatory responses, concluding that transparency and compliance obligations mitigate but do not resolve core legitimacy deficits. Ultimately, it argues that automated moderation faces normative limits and must remain subordinate to law, fairness, and contextual human judgment in speech governance.

"© 2026 by the Author(s). Published by IJAR under CC BY 4.0. Unrestricted use allowed with credit to the author."

Introduction:-

Classical free speech theory is grounded in a binary distinction between the state and the individual. Constitutional and human rights protections were developed to restrain government power, based on the assumption that censorship is primarily exercised by public authorities. This assumption no longer holds in the digital public sphere. Today, the most consequential decisions about speech are made not by governments or courts, but by private social media platforms operating at a global scale. Platforms such as Facebook, YouTube, Instagram, and X (formerly Twitter) regulate speech through privately drafted rules, enforced by a combination of human moderators and automated systems. These decisions affect billions of users and shape political discourse, cultural expression, and access to information worldwide. Yet, because platforms are legally categorized as private actors, their moderation practices largely fall outside the scope of constitutional free speech protections.

This chapter conceptualizes platforms as private governors of online speech, drawing primarily on Kate Klonick's seminal work. It argues that platforms have evolved into quasi-governing institutions that exercise regulatory power

comparable to states, while remaining insulated from the legal and democratic constraints traditionally imposed on public authorities. The chapter further examines how this governance role has given rise to quasi-legal moderation regimes, characterized by rulemaking, enforcement, and adjudication processes that resemble administrative law but lack its normative safeguards.

Platforms as Private Governors of Speech:-

Klonick's concept of platforms as "the new governors" provides a foundational framework for understanding contemporary content moderation (Klonick, 2018). She argues that platforms perform functions traditionally associated with sovereign governance, including norm creation, enforcement, dispute resolution, and sanctioning. Community standards operate as a form of internal law, defining acceptable and prohibited speech. Enforcement teams act as executive authorities, while appeal mechanisms and oversight bodies resemble judicial institutions. Unlike traditional intermediaries, platforms do not merely transmit speech. They actively curate, rank, suppress, and remove content, thereby shaping what users can see and say. This governance is not incidental but structural. Platform design choices, algorithmic ranking systems, and enforcement policies collectively determine the contours of permissible expression in the digital public sphere.

Crucially, this governance is exercised at scale. Decisions are applied globally, often without regard to local legal standards or cultural contexts. A single content policy drafted in California can determine the visibility of political speech in India, satire in France, or activism in authoritarian regimes. This transnational reach amplifies the regulatory power of platforms beyond that of many states.

Unlike public law, platform governance lacks democratic legitimacy. Community standards are drafted internally, often in opaque processes driven by corporate priorities such as brand safety, advertiser interests, and liability avoidance. Users have no meaningful role in shaping these rules, despite being subject to their enforcement. Klonick emphasizes that platforms borrow the language and structure of law without its accountability mechanisms (Klonick, 2018). Terms such as "appeals," "oversight," and "precedent" create the appearance of legality, while obscuring the fundamentally private nature of decision-making. This phenomenon raises concerns about legitimacy, fairness, and power asymmetry between platforms and users.

Absence of Constitutional Free Speech Protection:-

One of the most significant legal implications of platform governance is the absence of constitutional free speech protection. In most jurisdictions, constitutional rights apply only against the state, not private entities. Courts in the United States, the United Kingdom, and the European Union have consistently held that social media platforms are private actors, even when they host speech central to democratic discourse. In the U.S. context, the First Amendment restricts government censorship but does not constrain private platforms. Attempts to characterize platforms as public forums or state actors have largely failed. As a result, users whose content is removed or suppressed cannot invoke constitutional free speech rights, regardless of the social or political importance of their speech.

This doctrinal gap creates what Klonick describes as a free speech paradox (Klonick, 2018). Platforms exercise unprecedented control over speech but are subject to fewer legal constraints than governments. While states must justify restrictions under strict constitutional scrutiny, platforms can remove speech based on vague or discretionary standards, with minimal transparency or accountability. The privatization of speech regulation has profound implications for democratic theory. Public law traditionally balances competing interests such as security, equality, and expression through transparent processes and judicial review. Platform governance replaces this balancing exercise with corporate decision-making optimized for efficiency and risk management. Klonick (2020) notes that this privatization is not merely tolerated but often encouraged by governments. States increasingly rely on platforms to regulate harmful or illegal content, effectively outsourcing speech governance without extending corresponding constitutional obligations. This dynamic further entrenches the power of platforms while weakening public oversight.

Quasi-Legal Moderation Regimes:-

Platform moderation systems exhibit many features of legal regimes. They involve written rules, enforcement procedures, sanctions, and appeals. However, these systems lack key elements of the rule of law, including clarity, consistency, reason-giving, and independent adjudication. Klonick (2020) analyzes the creation of Facebook's Oversight Board as an attempt to address legitimacy deficits in platform governance. The Board was designed to

function as an independent adjudicatory body capable of reviewing content moderation decisions and issuing binding rulings. While innovative, the Oversight Board illustrates the limits of internal reform. It remains funded by the platform, constrained by its jurisdiction, and incapable of addressing the systemic incentives driving over-removal.

Quasi-legal moderation regimes suffer from significant procedural weaknesses. Rules are often vague, enforcement is inconsistent, and users receive limited explanations for decisions. Appeals mechanisms, where they exist, are frequently automated or under-resourced, offering little opportunity for meaningful redress. These weaknesses exacerbate power asymmetry between platforms and users. Platforms act as rulemakers, enforcers, and judges in their own systems, violating basic principles of separation of powers. Users, by contrast, lack bargaining power, legal remedies, or effective avenues for contestation.

Automation and the Intensification of Private Governance:-

While this chapter focuses on conceptual foundations, it is important to note that automation intensifies existing governance problems. Automated moderation systems enable platforms to enforce rules at scale, but they also entrench opacity and reduce opportunities for human judgment. Klonick (2018) warns that legitimacy in governance depends not only on outcomes but on processes perceived as fair and accountable. Automation undermines these perceptions by replacing reasoned judgment with probabilistic decision-making. When users cannot understand or challenge decisions, trust in platform governance erodes.

Implications for Free Speech Theory:-

The rise of platforms as private governors challenges core assumptions of free speech theory. Traditional frameworks focus on preventing state censorship, but offer limited tools for addressing private regulation at scale. This gap has prompted calls for new conceptual models that recognize platforms as hybrid institutions exercising public-like power. Klonick's work suggests that addressing this challenge requires more than importing constitutional norms into private law. It demands a rethinking of governance, accountability, and legitimacy in digital spaces. Whether through regulation, institutional design, or human rights frameworks, the unchecked power of private governors poses a fundamental challenge to freedom of expression in the twenty-first century. This chapter has established the conceptual foundation for analyzing automated content moderation by framing platforms as private governors of online speech. It has shown that platforms exercise regulatory power comparable to states, operate outside constitutional free speech constraints, and rely on quasi-legal moderation regimes that lack core rule-of-law safeguards. These dynamics create the conditions under which automated takedown systems can generate systemic harms to freedom of expression. Subsequent chapters build on this foundation by examining how automation exacerbates due process failures, entrenches bias, and challenges existing legal and human rights frameworks.

Automated Content Moderation as a Legal System:-

From Individual Decisions to Systems of Governance:-

Public debate around content moderation often frames platform decisions as isolated acts: a post removed, an account suspended, a video demonetized. This framing is misleading. As established in Chapter 1, platforms operate as private governors of online speech, exercising regulatory authority at scale. Automation transforms this governance from a series of discretionary judgments into an integrated system of rule application. Content moderation is therefore best understood not as a collection of individual decisions, but as a socio-technical legal system structured around risk management, scalability, and liability avoidance.

Douek (2022) argues that content moderation must be analyzed through a systems-thinking lens, emphasizing that harms do not arise merely from erroneous outcomes, but from the design, incentives, and feedback loops embedded within moderation architectures. Automated content moderation exemplifies this shift. Decisions are not made ex post through contextual deliberation, but ex ante through probabilistic classification models trained to predict rule violations at scale. This transformation has significant legal implications. When moderation is systematized, errors cease to be exceptional and instead become structural. Accountability, likewise, becomes diffuse and difficult to locate.

Automation enables platforms to apply content rules across billions of users and posts, but it also fundamentally alters the nature of decision-making. Instead of reasoned judgment grounded in interpretation, automated moderation relies on pattern recognition, statistical inference, and risk thresholds. These systems are designed to

prioritize speed, consistency, and volume, often at the expense of accuracy, contextual sensitivity, and procedural fairness. As a result, automated moderation systems function less like discretionary enforcement mechanisms and more like regulatory infrastructures that predefine the boundaries of permissible speech.

This chapter conceptualizes automated content moderation as a legal system in its own right. It argues that automation restructures platform governance around three core dynamics: system-level enforcement rather than case-by-case adjudication, risk-based decision-making driven by liability exposure, and the emergence of structural harms that cannot be remedied through individual appeals or error correction. These dynamics distinguish automated moderation from both traditional censorship and human-mediated content regulation.

The Architecture of Automated Moderation:-

Automated content moderation systems operate through layered technical architectures that combine machine learning models, rule-based filters, and human oversight. At the front end, user-generated content is scanned using classifiers trained on large datasets to detect prohibited material, such as hate speech, copyright infringement, or misinformation. These classifiers assign probability scores indicating the likelihood that content violates platform rules. Content exceeding predefined thresholds may be removed automatically, deprioritized algorithmically, or flagged for human review.

The European Parliament (2020) report emphasizes that these systems are not neutral tools but reflect design choices shaped by regulatory pressure, commercial incentives, and technical constraints. Decisions about training data, classification thresholds, and model optimization determine which types of speech are most likely to be targeted. For example, systems optimized to minimize false negatives may aggressively over-remove borderline content, while systems tuned to reduce false positives may allow harmful speech to persist. These trade-offs are rarely transparent to users and are often calibrated to serve platform interests rather than expressive rights.

Automation also operates across multiple stages of content circulation. Moderation does not begin only after content is posted. Increasingly, platforms deploy upload filters, pre-publication checks, and real-time ranking adjustments that shape visibility before speech reaches an audience. This preemptive governance transforms moderation from reactive enforcement into proactive control. Speech is constrained not through explicit prohibition, but through infrastructural design that determines what can be seen, shared, or monetized.

Crucially, automated moderation systems are embedded within broader platform ecosystems. Content decisions interact with recommendation algorithms, advertising systems, and user engagement metrics. A single classification error can cascade through the system, affecting reach, reputation, and economic opportunity. As Douek (2022) notes, moderation systems must therefore be evaluated holistically, not as isolated enforcement tools but as components of a complex regulatory environment.

Moderation as Risk Management and Liability Avoidance:-

Automation is not adopted primarily to improve the quality of speech governance. Rather, it functions as a risk-management mechanism designed to reduce legal, financial, and reputational exposure. Platforms face increasing pressure from governments to address illegal and harmful content, including terrorism, copyright infringement, and disinformation. At the same time, they must satisfy advertisers, investors, and users who demand “safe” online environments. Automated moderation offers a scalable solution to these competing demands.

Risk-based enforcement is central to this logic. Instead of asking whether content is lawful or socially valuable, automated systems ask whether content presents an unacceptable level of risk. This risk may relate to potential regulatory penalties, advertiser backlash, or public controversy. As a result, moderation decisions are shaped less by normative judgments about expression and more by cost-benefit calculations. Speech that is ambiguous, controversial, or context-dependent is disproportionately affected, as it carries higher uncertainty and therefore higher perceived risk.

The European Parliament (2020) report highlights how liability regimes incentivize over-removal. When platforms are held responsible for failing to remove unlawful content, but face minimal consequences for removing lawful speech, automation naturally skews toward caution. This asymmetry is exacerbated by tight compliance deadlines and large-scale content volumes, which leave little room for contextual assessment. Automated systems are therefore designed to err on the side of removal, even when this results in lawful speech being suppressed.

This dynamic reinforces the privatization of speech regulation identified in Chapter 1. Platforms are not merely implementing legal requirements; they are translating them into internal risk frameworks that prioritize corporate survival over expressive freedom. Automation becomes the mechanism through which legal uncertainty is resolved in favor of control. The resulting system bears little resemblance to public law enforcement, which typically requires proportionality, justification, and individualized assessment.

The Legal Meaning of Automation:-

Understanding automated moderation as a legal system requires moving beyond technical descriptions to examine its normative implications. Automation alters the relationship between rules, decision-makers, and subjects of regulation. In traditional legal systems, rules are interpreted by human agents who are accountable through institutional mechanisms such as judicial review. Automated systems, by contrast, apply rules through statistical proxies, without interpretation or justification in the conventional sense.

Douek (2022) argues that this shift challenges foundational legal concepts, including notice, explanation, and responsibility. When a moderation decision is made by an algorithm, it is often unclear which rule was applied, how it was interpreted, or why a particular outcome was reached. Users may receive generic notifications citing broad policy categories, but these explanations lack the specificity required for meaningful contestation. The absence of reason-giving undermines the legitimacy of moderation decisions and weakens procedural safeguards.

Automation also diffuses responsibility. Decisions emerge from interactions between data, models, thresholds, and organizational policies. No single actor can be easily identified as the decision-maker. This diffusion complicates accountability, as legal remedies typically require attributing responsibility to a human or institutional agent. Platforms may attribute errors to technical limitations, while regulators struggle to assess compliance in the absence of transparency.

From a legal perspective, automated moderation blurs the boundary between rule application and rule creation. When systems are trained on past enforcement data, they internalize prior biases and enforcement patterns. Over time, these patterns become entrenched, effectively generating new norms without formal rulemaking. This process resembles common law development, but without public reasoning, precedent, or oversight. Automation thus becomes a source of norm production as well as enforcement.

Structural Harms of Automated Moderation:-

The harms associated with automated moderation are often framed as technical errors or accuracy problems. This framing is inadequate. While false positives and false negatives are significant, they represent symptoms rather than causes. The deeper problem lies in the structural design of automated systems, which systematically disadvantage certain forms of speech and certain groups of speakers.

One structural harm is the erosion of contextual judgment. Automated systems struggle to interpret satire, irony, political dissent, and culturally specific expression. These forms of speech rely on context that cannot be easily captured through pattern recognition. As a result, they are more likely to be misclassified and removed. This effect is magnified for marginalized communities whose linguistic practices differ from dominant training data.

Another structural harm is the chilling effect produced by opacity and unpredictability. When users cannot anticipate how rules will be applied, they may self-censor to avoid sanctions. The European Parliament (2020) report notes that automated moderation can discourage participation in public discourse, particularly for activists, journalists, and minority voices. This chilling effect undermines the democratic potential of digital platforms, even in the absence of overt censorship.

Structural harms are also resistant to correction. Individual appeals may restore specific content, but they do not address the underlying system design that produced the error. As long as incentives favor speed and risk minimization, automated systems will continue to generate similar outcomes. This persistence distinguishes automated moderation from isolated enforcement mistakes and underscores the need for systemic analysis.

Automation and the Illusion of Neutrality:-

Automation is often justified through claims of objectivity and consistency. Algorithms are portrayed as impartial tools that apply rules uniformly, free from human bias. This narrative is misleading. Automated systems reflect the

values, assumptions, and priorities embedded in their design. Choices about what to measure, how to classify, and where to set thresholds are inherently normative.

Douek (2022) emphasizes that system design choices determine whose speech is protected and whose is suppressed. For example, systems optimized to detect hate speech may disproportionately target reclaimed slurs or political critique, while failing to capture coded language used by extremist groups. These outcomes are not accidental but flow from design trade-offs that prioritize efficiency over deliberation.

The illusion of neutrality also serves to legitimize platform power. By attributing decisions to algorithms, platforms can deflect criticism and obscure the human and institutional choices underlying moderation practices. Automation thus functions as both a technical and rhetorical tool, masking governance decisions behind claims of inevitability and scale.

This chapter has argued that automated content moderation operates as a legal system rather than a collection of isolated technical interventions. Automation restructures platform governance around system-level enforcement, risk-based decision-making, and structural harms that cannot be addressed through individual remedies. Drawing on Douek's systems-thinking framework and the European Parliament's analysis of algorithmic filtering, the chapter has shown that automation transforms the nature of speech regulation, amplifying private power while weakening traditional legal safeguards.

By reframing automated moderation as a form of legal ordering, this chapter lays the groundwork for the subsequent analysis of due process failures, algorithmic arbitrariness, and human rights constraints. The next chapter builds on this foundation by examining how automated systems undermine procedural fairness through opaque notice mechanisms and ineffective appeals, further entrenching the legitimacy crisis of private speech governance.

Due Process Failures in Automated Takedown Systems:-

Having established that automated content moderation functions as a legal system rather than a series of isolated enforcement decisions, this chapter examines the procedural consequences of that transformation. In public law, the legitimacy of regulatory authority depends not only on substantive outcomes but on the fairness of procedures through which decisions are made. Core due process guarantees such as notice, explanation, and the right to appeal serve to constrain power, enable contestation, and protect individuals against arbitrary decision-making.

In the context of platform governance, these procedural safeguards are largely absent or functionally hollow. Automated takedown systems routinely remove or restrict content without providing meaningful notice, intelligible explanations, or effective avenues for appeal. While platforms often claim to offer procedural protections through notifications and review mechanisms, empirical evidence suggests that these systems fail to meet even minimal standards of procedural fairness.

This chapter argues that due process failures in automated moderation are not accidental or remediable through incremental improvements. Rather, they are structural features of a system optimized for scale, speed, and risk management. Drawing on Are's (2024) empirical analysis of appeals systems and Douek's (2019) critique of platform adjudicatory institutions, the chapter demonstrates how automation undermines procedural justice, entrenches power asymmetries, and exacerbates the legitimacy crisis of private speech governance.

Procedural Fairness and the Problem of Automation:-

Procedural fairness rests on several foundational principles: individuals must be informed of decisions affecting their rights or interests, they must be given reasons for those decisions, and they must have a meaningful opportunity to challenge them before an impartial decision-maker. These principles are central to administrative law and human rights jurisprudence, reflecting the idea that power must be exercised transparently and accountably.

Automated moderation systems depart from these principles in fundamental ways. Decisions are generated through probabilistic models that classify content based on patterns rather than interpretation. As discussed in Chapter 2, these systems operate *ex ante*, applying generalized rules across vast quantities of content. This design makes individualized reasoning costly and, in many cases, technically infeasible.

The result is a procedural environment in which users are subject to consequential decisions without access to the information necessary to understand or contest them. Automation replaces reason-giving with rule invocation, explanation with categorization, and adjudication with threshold-based filtering. Due process is thus reduced to a formal gesture rather than a substantive safeguard.

Importantly, these failures cannot be fully understood by focusing on individual errors or system malfunctions. They arise from the logic of automation itself. When moderation is conceived as a systems problem rather than a legal one, procedural fairness becomes a secondary concern, subordinated to efficiency and compliance metrics.

Opaque Notice Systems:-

Notice is the first procedural step through which regulated subjects become aware of enforcement actions. In legal systems, notice must be timely, specific, and sufficiently detailed to enable understanding and response. In automated content moderation, notice systems rarely meet these criteria. When content is removed or restricted, users typically receive standardized notifications referencing broad policy categories, such as “hate speech,” “misinformation,” or “copyright infringement.” These notices often lack specificity regarding which aspect of the content violated platform rules, how the rule was interpreted, or whether automation or human review was involved. As a result, users are left to infer the reasons for enforcement from vague descriptions that provide little practical guidance.

Are (2024) characterizes this phenomenon as a core failure of algorithmic justice. Based on qualitative analysis of user experiences, Are demonstrates that platform notices frequently fail to communicate actionable information. Users report confusion, frustration, and uncertainty about how to comply with rules in the future. This opacity undermines both fairness and predictability, two key components of legitimate governance. The problem is exacerbated by automation’s reliance on probabilistic classification. Because decisions are based on likelihood scores rather than determinate findings, platforms often avoid disclosing specific rationales. Doing so would expose uncertainty, reveal enforcement thresholds, or invite strategic behavior by users. As a result, platforms opt for generic notices that obscure the actual basis of decision-making. From a legal perspective, opaque notice systems undermine the right to be heard. Without knowing why content was removed, users cannot meaningfully challenge decisions or modify future behavior. Notice becomes a one-way communication that informs users of outcomes without enabling participation in the decision-making process.

Explanation and the Absence of Reason-Giving:-

Explanation is closely linked to notice but serves a distinct function. While notice informs individuals that a decision has been made, explanation provides the reasons behind that decision. In administrative law, reason-giving is essential for accountability, enabling both affected individuals and reviewing bodies to assess whether decisions were lawful, proportionate, and consistent. Automated moderation systems largely dispense with reason-giving. Users may be told that content violated a particular policy, but they are rarely informed how that policy was applied or which elements of their content triggered enforcement. This absence of explanation is not merely a design flaw; it reflects the epistemic limits of machine learning systems.

Because automated classifiers operate through statistical inference rather than rule-based reasoning, they do not generate explanations in a form that maps easily onto legal categories. Outputs consist of confidence scores and feature correlations, not articulated justifications. Translating these outputs into human-understandable reasons would require additional interpretive layers, increasing complexity and cost. Platforms have largely chosen not to invest in such interpretive infrastructure. As Douek (2019) observes, explanation is treated as an optional feature rather than a foundational requirement of legitimate governance. This choice reflects institutional priorities rather than technical impossibility. Platforms prioritize throughput and compliance over procedural transparency. The absence of explanation has significant normative consequences. Without reasons, decisions appear arbitrary, even when they are statistically consistent. Users cannot distinguish between justified enforcement and error, nor can they evaluate whether rules are being applied evenly. Explanation is replaced by trust in the system, a trust that is frequently unwarranted and difficult to sustain.

Appeals as Formality Rather Than Remedy:-

Appeals mechanisms are often presented by platforms as evidence of procedural fairness. In practice, however, these mechanisms frequently fail to provide meaningful redress. Are (2024) provides a detailed empirical account of what she terms “dysfunctional appeals,” highlighting systemic deficiencies in design, accessibility, and effectiveness.

One central problem is the automation of appeals themselves. In many cases, appeals are processed through the same or similar automated systems that generated the original decision. This creates a closed feedback loop in which errors are reproduced rather than corrected. Users may submit additional context or clarification, but the system lacks the capacity to meaningfully interpret such input. Even when human review is involved, appeals processes are often under-resourced and constrained by strict time limits. Reviewers may rely on internal guidelines that prioritize consistency over deliberation, leading to rubber-stamping of initial decisions. The scale of content moderation makes individualized reconsideration impractical, reinforcing a bias toward affirming takedowns.

Are's findings indicate that users frequently perceive appeals as performative rather than substantive. Many report receiving identical responses to appeals, suggesting minimal engagement with the merits of their cases. This perception erodes confidence in platform governance and discourages future contestation. From a procedural standpoint, ineffective appeals mechanisms fail to satisfy basic fairness requirements. A right to appeal that does not offer a realistic chance of reversal is functionally meaningless. Appeals become a symbolic gesture that legitimizes enforcement without constraining it.

Power Asymmetry and the Illusion of Choice:-

Due process failures in automated moderation must be understood within a broader context of power asymmetry. Platforms occupy a dominant position, controlling the infrastructure through which speech is disseminated and monetized. Users, by contrast, lack bargaining power, legal recourse, or alternative venues for expression of comparable reach. This asymmetry is compounded by the contractual framing of platform governance. Terms of service typically reserve broad discretion for platforms to remove content and suspend accounts, while limiting liability and dispute resolution options for users. Procedural protections are framed as voluntary features rather than enforceable rights.

Douek (2019) critiques this dynamic in her analysis of Facebook's Oversight Board, noting that internal review mechanisms cannot fully compensate for the absence of external accountability. While the Board represents an innovative attempt to introduce adjudicatory norms into platform governance, it operates within constraints imposed by the platform itself. Its jurisdiction is limited, its case selection discretionary, and its authority dependent on corporate compliance. The existence of such bodies may create the appearance of procedural fairness without addressing underlying power imbalances. Users are offered the illusion of choice and participation, while platforms retain ultimate control over rules, enforcement, and remedies. Due process becomes a branding exercise rather than a substantive constraint on power.

Automation, Scale, and the Erosion of Individualized Justice:-

A recurring justification for procedural shortcuts in content moderation is scale. Platforms argue that individualized notice, explanation, and appeal are infeasible given the volume of content processed daily. Automation is presented as the only viable solution to managing speech at scale. This argument, while descriptively accurate, has normative implications. It implies that procedural justice is incompatible with large-scale governance, and that individual rights must be sacrificed for efficiency. Such a position departs sharply from legal traditions that treat due process as a non-negotiable constraint, even in contexts involving mass administration. Automated moderation prioritizes uniformity over responsiveness, speed over deliberation, and predictability over contestability. These priorities may be defensible from a business perspective, but they undermine the legitimacy of governance over expressive activity. When individuals are subjected to sanctions without meaningful procedural protections, governance becomes coercive rather than consensual.

Are (2024) emphasizes that procedural failures disproportionately affect marginalized users, whose speech is more likely to be misclassified and less likely to be successfully appealed. Automation thus exacerbates existing inequalities, reinforcing dominant norms while suppressing dissenting or unconventional expression. This chapter has shown that automated takedown systems systematically fail to provide procedural fairness. Opaque notice systems, the absence of reason-giving, and ineffective appeals mechanisms deprive users of meaningful opportunities to understand and challenge moderation decisions. These failures are not incidental but flow from the structural logic of automation, which prioritizes scale and risk management over individualized justice. By undermining due process, automated moderation intensifies the legitimacy deficit identified in Chapters 1 and 2. Platforms exercise regulatory authority over speech without adhering to the procedural norms that justify such power in public law. Appeals mechanisms and oversight bodies, while symbolically important, do little to correct systemic imbalances. The next chapter builds on this analysis by examining how automation not only undermines

procedure but also produces substantively arbitrary outcomes, particularly in the context of copyright enforcement and prototype censorship regimes. Together, these chapters demonstrate that the automation of speech governance poses fundamental challenges to both fairness and freedom of expression.

Algorithmic Arbitrariness and Copyright-Based Automation:-

From Procedural Failure to Substantive Arbitrariness:-

The preceding chapter demonstrated that automated takedown systems systematically fail to provide procedural fairness through meaningful notice, explanation, and appeal. This chapter advances the analysis by examining the substantive consequences of those failures. When procedural safeguards erode, the problem is not merely one of unfair process, but of arbitrary outcomes. Automation does not simply make moderation faster or more efficient; it reshapes the substance of speech regulation by producing inconsistent, biased, and often inexplicable enforcement decisions. This chapter argues that algorithmic arbitrariness is a defining feature of automated content moderation, not a technical anomaly. Drawing on Gomez et al. (2024), it shows that automated systems generate unpredictable and uneven outcomes that disproportionately affect marginalized speakers and context-dependent forms of expression. These outcomes are reinforced by legal frameworks that incentivize over-removal, most notably copyright law under the Digital Millennium Copyright Act (DMCA).

Copyright enforcement represents the earliest and most entrenched form of automated speech regulation. As Carpou (2015) and Urban and Quilter (2011) demonstrate, copyright takedown systems pioneered the use of automated notices, strict liability incentives, and minimal procedural safeguards. This chapter contends that copyright-based automation has functioned as a prototype censorship regime, providing both the technical infrastructure and the normative justification for the expansion of automation into general speech governance. What began as a narrowly framed intellectual property enforcement mechanism has become a model for platform-wide control over expression.

Algorithmic Arbitrariness as a Structural Phenomenon:-

Arbitrariness in legal decision-making is traditionally understood as the absence of reasoned justification, consistency, or principled application of rules. In public law, arbitrariness is constrained through doctrines such as proportionality, reason-giving, and judicial review. In automated content moderation, these constraints are largely absent, allowing arbitrariness to emerge at scale. Gomez et al. (2024) provide a detailed empirical analysis of algorithmic arbitrariness in content moderation, demonstrating that automated systems frequently produce inconsistent outcomes across similar content. Identical or near-identical posts may be treated differently depending on context, timing, language, or user characteristics. These inconsistencies are not random errors but reflect the probabilistic nature of machine-learning systems trained on incomplete and biased datasets.

Algorithmic arbitrariness arises from several interrelated factors. First, classifiers rely on statistical correlations rather than semantic understanding. They detect patterns associated with rule violations but cannot reliably interpret meaning, intent, or context. Second, enforcement thresholds are often adjusted dynamically in response to external pressures, such as regulatory scrutiny or public controversy. These adjustments alter enforcement behavior without transparent justification, further undermining consistency. Third, moderation systems operate within complex platform ecosystems, where visibility, ranking, and monetization decisions intersect with takedown mechanisms. The result is a form of governance in which outcomes are neither predictable nor intelligible to users. Unlike human adjudication, where inconsistency may be challenged through appeals or precedent, algorithmic arbitrariness is embedded in system design. It cannot be fully corrected without rethinking the underlying architecture of automation.

Bias and Disparate Impact in Automated Moderation:-

Algorithmic arbitrariness is closely linked to bias. While platforms often frame automation as neutral and objective, empirical research consistently demonstrates disparate impacts across different groups and forms of expression. Gomez et al. (2024) show that automated moderation systems disproportionately misclassify content produced by marginalized communities, particularly when language use deviates from dominant norms. Bias enters automated systems through training data, which reflects historical enforcement patterns and societal inequalities. Content created by minority groups is often over-represented in datasets labeled as problematic, leading classifiers to associate certain linguistic or cultural markers with rule violations. This dynamic results in higher false-positive rates for already marginalized speakers, reinforcing existing power asymmetries in the digital public sphere. Bias is also exacerbated by the risk-based enforcement logic discussed in Chapter 2. Content that is politically contentious,

culturally specific, or context-dependent is treated as higher risk due to uncertainty. Automated systems therefore err on the side of removal, suppressing speech that does not conform to easily classifiable categories. This effect is particularly pronounced in discussions of race, gender, religion, and political dissent. Importantly, biased outcomes are difficult to detect and contest. As shown in Chapter 3, users rarely receive explanations that would allow them to identify discriminatory patterns. Appeals mechanisms, where they exist, focus on individual cases rather than systemic disparities. Bias thus operates invisibly, masked by claims of technical inevitability and scale.

Copyright Enforcement as the Prototype of Automation:-

Copyright law provides the clearest illustration of how automated moderation produces arbitrary outcomes. Long before platforms began automating hate speech or misinformation enforcement, they relied on automated copyright takedown systems to comply with the DMCA's notice-and-takedown regime. These systems were designed to process vast volumes of infringement claims quickly, with minimal human intervention. Carpou (2015) traces the rise of automated copyright enforcement, showing how technological tools such as content identification systems enabled platforms to manage liability risks associated with user-generated content. These systems detect copyrighted material through pattern matching, often without regard to context or lawful exceptions such as fair use. As a result, lawful speech is routinely removed or blocked based on automated determinations.

Urban and Quilter (2011) document the chilling effects of DMCA takedown notices, demonstrating that copyright enforcement frequently suppresses non-infringing content, including political speech, parody, and criticism. Automated processing exacerbates these effects by prioritizing speed and volume over accuracy. Platforms have little incentive to scrutinize claims, as liability attaches to failure to remove content, not to over-removal. Copyright automation thus established a governance model characterized by strict liability incentives, minimal procedural safeguards, and high error tolerance. This model normalized the idea that lawful speech could be sacrificed for efficiency and compliance. Over time, it became an accepted cost of operating large-scale platforms.

DMCA Logic and the Normalization of Over-Removal:-

The DMCA's notice-and-takedown framework incentivizes platforms to remove content upon receiving a claim, regardless of its validity. Urban and Quilter (2011) show that this asymmetry leads to systematic over-removal, as platforms face greater risks from retaining contested content than from removing it. Automated systems amplify this tendency by eliminating human discretion at early stages of enforcement. This liability-driven logic has broader implications beyond copyright. As platforms adopt similar approaches to other categories of content, including terrorism and misinformation, the DMCA model becomes a template for automated speech regulation. The assumption that rapid removal is preferable to careful assessment migrates from intellectual property law into general content governance. Copyright enforcement also illustrates how automation obscures accountability. When content is removed due to an automated match, responsibility is diffused between rights holders, platforms, and algorithms. Users often lack clarity about who initiated the takedown or how to challenge it. This diffusion mirrors the accountability gaps identified in Chapters 2 and 3, reinforcing the privatization of speech regulation.

Expansion from Intellectual Property to General Speech Control:-

The expansion of automation from copyright enforcement to broader speech regulation represents a significant shift in the governance of online expression. Copyright law, while controversial, is at least grounded in a defined legal framework with established exceptions and limitations. When its enforcement mechanisms are repurposed for other forms of speech, those legal constraints are often absent. Platforms increasingly deploy automated systems to address categories such as hate speech, extremism, and misinformation. Unlike copyright, these categories lack clear statutory definitions and involve complex normative judgments. Applying automation to such domains magnifies arbitrariness, as classifiers struggle to capture context, intent, and social meaning.

The normalization of copyright automation has also shaped user expectations and regulatory discourse. Over-removal is treated as an unfortunate but acceptable trade-off, rather than a systemic failure. This mindset discourages the development of robust safeguards and legitimizes the expansion of automation into areas where expressive stakes are higher. From a legal perspective, this expansion raises serious concerns. It extends a liability-driven enforcement model into domains central to democratic discourse, without corresponding protections for freedom of expression. Automated moderation thus becomes a tool of generalized speech control, exercised by private actors without public accountability.

Algorithmic Arbitrariness and the Rule of Law:-

Algorithmic arbitrariness poses a direct challenge to rule-of-law principles. Legal systems aspire to govern through general, predictable, and publicly accessible rules applied consistently to similarly situated individuals. Automated moderation systems fall short of these ideals. Rules are applied through opaque models, outcomes vary unpredictably, and users lack access to meaningful remedies. Gomez et al. (2024) emphasize that arbitrariness is not merely a matter of accuracy but of legitimacy. Even perfectly accurate systems would raise concerns if they operate without transparency or accountability. In practice, automated moderation systems are neither accurate nor accountable, compounding legitimacy deficits. The rule of law also requires that power be constrained through institutional checks. In automated moderation, platforms act as rulemakers, enforcers, and adjudicators, as discussed in Chapter 1. Automation intensifies this concentration of power by reducing opportunities for human intervention and external scrutiny. Arbitrariness becomes systemic rather than episodic.

This chapter has argued that algorithmic arbitrariness is an inherent feature of automated content moderation, reinforced by legal and technical frameworks that prioritize efficiency and liability avoidance over expressive freedom. Drawing on empirical research and copyright law scholarship, it has shown how automated systems produce inconsistent and biased outcomes, particularly for context-dependent and marginalized speech. Copyright-based automation under the DMCA has served as a prototype for broader speech regulation, normalizing over-removal and procedural minimalism. As this model expands beyond intellectual property into general content governance, it amplifies the risks to freedom of expression identified in earlier chapters. Together with the procedural failures examined in Chapter 3, algorithmic arbitrariness underscores the inadequacy of current moderation systems as mechanisms of legitimate governance. The next chapter builds on this analysis by examining how international human rights law seeks to constrain automated moderation through principles of legality, necessity, proportionality, and procedural safeguards, and why these constraints remain difficult to enforce in practice.

Human Rights Constraints on Automated Moderation:-**Human Rights as the Last External Constraint:-**

The preceding chapters have demonstrated that automated content moderation systems operate as private regulatory regimes marked by procedural failure and substantive arbitrariness. Platforms exercise extensive control over speech through opaque, risk-driven, and biased systems that lack meaningful accountability. Against this backdrop, international human rights law appears as one of the few normative frameworks capable of imposing external constraints on private speech governance. Human rights law, particularly freedom of expression under Article 19 of the International Covenant on Civil and Political Rights (ICCPR), offers a set of standards designed to limit arbitrary interference with expression. These standards include legality, legitimate aim, necessity, proportionality, and procedural safeguards. In theory, they provide a robust framework for evaluating restrictions on speech, regardless of whether those restrictions originate from public or private actors.

In practice, however, the application of human rights norms to automated content moderation is deeply contested. Platforms are private entities, states increasingly outsource enforcement to them, and automation introduces forms of restriction that are diffuse, preemptive, and difficult to attribute. This chapter examines how human rights law seeks to constrain automated moderation, with a particular focus on upload filters and EU regulatory frameworks, and why those constraints remain fragile and incomplete. Drawing on Romero-Moreno (2020) and Penagos (2025), the chapter argues that automated moderation systems frequently fail to satisfy core Article 19 requirements, particularly proportionality and procedural safeguards. While human rights discourse has increasingly influenced regulatory design, especially in the European Union, automation continues to undermine the practical effectiveness of these protections.

Article 19 ICCPR and Freedom of Expression Standards:-

Article 19 of the ICCPR guarantees the right to freedom of expression, including the freedom to seek, receive, and impart information and ideas of all kinds. Any restriction on this right must meet a strict three-part test: it must be provided by law, pursue a legitimate aim, and be necessary and proportionate to that aim. These standards are designed to prevent arbitrary interference with expression and to ensure that restrictions are justified, transparent, and narrowly tailored. In traditional contexts, they apply to state action, with courts and human rights bodies assessing whether laws and enforcement practices comply with these criteria. The relevance of Article 19 to platform moderation arises from two developments. First, platforms have become central venues for public discourse, performing functions that resemble public forums. Second, states increasingly rely on platforms to

enforce speech restrictions, effectively delegating regulatory authority without extending corresponding constitutional or human rights obligations. This delegation creates a regulatory gap in which private actors exercise public-like power without being directly bound by human rights law. Human rights bodies have responded by emphasizing states' positive obligations to protect freedom of expression against interference by private actors. This includes ensuring that regulatory frameworks governing platforms are compatible with Article 19 standards. However, translating these abstract obligations into concrete constraints on automated moderation has proven challenging.

Legality and the Problem of Vague Rules:-

The requirement of legality demands that restrictions on expression be prescribed by law in a manner that is accessible, clear, and foreseeable. Individuals must be able to understand what conduct is prohibited and anticipate the consequences of their actions. Automated moderation systems struggle to satisfy this requirement. Platform rules are often drafted in broad, indeterminate language and applied through machine-learning models that operate probabilistically rather than deterministically. As a result, users cannot reliably predict how their speech will be treated. Enforcement thresholds may change dynamically, and internal policies may be updated without notice.

Romero-Moreno (2020) notes that upload filters mandated or encouraged by EU copyright law exacerbate this problem. Filters operate through pattern matching and similarity detection, identifying content that resembles protected material. These systems lack the capacity to interpret legal exceptions such as quotation, parody, or criticism, which are central to copyright law's balance between protection and expression. As a result, lawful expression is routinely restricted without clear legal justification. From a human rights perspective, this lack of foreseeability undermines the legality of automated restrictions. When speech is filtered or removed based on opaque technical criteria rather than publicly articulated legal standards, the rule of law gives way to rule by system design.

Proportionality and the Disproportionate Impact of Upload Filters:-

Proportionality is a core principle of human rights law, requiring that restrictions on expression be appropriate to achieve a legitimate aim and not excessive in relation to that aim. Measures must be the least restrictive means available and must strike a fair balance between competing interests. Upload filters pose a particular challenge to proportionality. By design, they operate *ex ante*, preventing content from being published before any contextual assessment can occur. This preemptive approach treats all potentially infringing content as equally suspect, regardless of its actual legal status or expressive value. Romero-Moreno (2020) argues that such systems are inherently disproportionate. They impose sweeping restrictions on expression in order to prevent a subset of unlawful uses, resulting in widespread collateral suppression of lawful speech. This effect is especially pronounced for creative, political, and transformative works that rely on quotation or reuse.

Penagos (2025) extends this critique by examining the implementation of upload filters under EU law, highlighting the tension between copyright enforcement and fundamental rights. While EU legislation formally acknowledges the need to protect freedom of expression, the technical reality of filtering systems makes nuanced balancing practically unattainable. Automated systems cannot reliably distinguish between infringing and non-infringing uses, leading to systematic over-blocking. The disproportionate nature of upload filters is not an unintended side effect but a predictable consequence of automation. When systems are optimized to avoid false negatives, they inevitably increase false positives. From a human rights perspective, this trade-off is deeply problematic, as it prioritizes enforcement efficiency over expressive freedom.

Procedural Safeguards as Human Rights:-

Procedural safeguards are not merely administrative conveniences; they are integral to the protection of human rights. Article 19 jurisprudence emphasizes the importance of safeguards such as notice, explanation, and the right to an effective remedy. These safeguards ensure that restrictions on expression can be challenged and reviewed, preventing abuse of power. As demonstrated in Chapter 3, automated moderation systems routinely fail to provide meaningful procedural protections. Notices are vague, explanations are absent or unintelligible, and appeals mechanisms are ineffective. From a human rights perspective, these failures amount to violations of procedural guarantees inherent in the right to freedom of expression. Penagos (2025) highlights the importance of embedding procedural safeguards into the design of automated systems. This includes requirements for transparency, human oversight, and accessible remedies. However, the implementation of such safeguards faces significant resistance from platforms, which view them as costly and incompatible with scale. Moreover, procedural safeguards are often treated as secondary to substantive compliance. Platforms may focus on meeting formal transparency obligations

while neglecting the quality and effectiveness of remedies. This formalistic approach undermines the spirit of human rights protections, reducing them to box-ticking exercises.

Attribution, State Responsibility, and the Problem of Private Enforcement:-

A central challenge in applying human rights law to automated moderation is attribution. Human rights obligations traditionally apply to states, not private entities. When platforms restrict speech through automated systems, determining responsibility becomes complex. States often argue that moderation decisions are made by private actors and therefore fall outside their direct control. Platforms, in turn, claim that they are merely complying with legal obligations or acting to protect users and advertisers. This mutual deflection creates a responsibility vacuum in which no actor fully accepts accountability for rights violations.

Human rights bodies have attempted to address this gap by emphasizing states' positive obligations. When states mandate or incentivize the use of upload filters, they bear responsibility for ensuring that resulting restrictions comply with Article 19 standards. However, enforcement of these obligations is weak, particularly at the international level. Automation further complicates attribution by diffusing decision-making across technical systems. Restrictions emerge from the interaction of algorithms, datasets, and policy choices, making it difficult to identify a single point of responsibility. This diffusion undermines traditional accountability mechanisms and weakens the practical force of human rights law.

Human Rights Law and Its Structural Limits:-

While human rights law provides a powerful normative framework, it faces structural limitations in addressing automated moderation. Its reliance on state responsibility, individualized harm, and post hoc review makes it ill-suited to addressing systemic, large-scale restrictions imposed by private platforms. Human rights litigation is slow, reactive, and resource-intensive. It is poorly equipped to respond to the volume and speed of automated takedowns. Even when violations are recognized, remedies are often symbolic, failing to address underlying system design. Romero-Moreno (2020) cautions against viewing human rights law as a panacea. Without complementary regulatory and institutional reforms, human rights standards risk being absorbed into the logic of automation, providing legitimacy without constraint. References to proportionality and safeguards may appear in legislation, while actual enforcement practices continue to prioritize efficiency and risk management.

This chapter has examined the capacity of human rights law to constrain automated content moderation. While Article 19 ICCPR provides clear normative standards for protecting freedom of expression, automated systems frequently fail to meet requirements of legality, proportionality, and procedural fairness. Upload filters, in particular, exemplify the disproportionate and preemptive nature of automated restrictions on speech. Human rights law remains an essential reference point, offering a language of critique and a set of principles against which automation can be evaluated. However, its effectiveness is limited by structural factors, including the privatization of enforcement, the diffusion of responsibility, and the scale of automated governance. The next chapter turns to regulatory responses, focusing on the European Union's attempts to operationalize human rights principles through transparency and compliance obligations. It examines whether these experiments meaningfully constrain private automation or merely formalize its dominance.

Regulatory Responses and Legal Experimentation:-

Regulation After Automation:-

The preceding chapters have demonstrated that automated content moderation functions as a private system of speech governance marked by procedural failure, substantive arbitrariness, and weak human rights safeguards. In response to growing concern about the power of digital platforms, governments have increasingly turned to regulation as a means of reasserting public control. Among these efforts, the European Union has emerged as the most active regulatory actor, experimenting with legal frameworks intended to constrain platform behavior without dismantling platform infrastructure. This chapter examines the European Union's regulatory approach to automated content moderation, focusing on transparency obligations, systemic risk assessments, and compliance-based governance. It evaluates whether these measures meaningfully address the legitimacy deficits identified in earlier chapters or whether they instead normalize automation by embedding it within formal regulatory processes.

Drawing on the European Parliament (2020) report and Douek's (2022) systems-thinking analysis, the chapter argues that EU regulation functions less as a substantive constraint on automated moderation than as a procedural overlay. While regulatory initiatives improve visibility and documentation, they leave core structural incentives

intact. Automation remains driven by scale, risk management, and liability avoidance, with regulation serving primarily to audit rather than alter these dynamics.

The European Union as a Regulatory Laboratory:-

The European Union has positioned itself as a global leader in digital platform regulation. Through initiatives such as the Digital Services Act (DSA) and related instruments, the EU seeks to impose obligations on platforms to address illegal content, systemic risks, and harms to fundamental rights. Unlike the largely laissez-faire approach historically taken in other jurisdictions, EU regulation emphasizes precaution, oversight, and harmonization across member states. The European Parliament (2020) characterizes the EU's approach as experimental, reflecting an attempt to balance competing objectives. On the one hand, the EU aims to protect freedom of expression and fundamental rights. On the other, it seeks to address harms associated with disinformation, hate speech, and copyright infringement. Automation is treated as both a problem and a solution: a source of risk that must be managed, and a necessary tool for enforcement at scale.

This dual framing shapes regulatory design. Rather than prohibiting automated moderation, EU law assumes its inevitability and focuses on procedural constraints. Platforms are required to assess risks, document systems, and report on enforcement practices. The underlying assumption is that transparency and accountability mechanisms can correct excesses without disrupting platform operations. This experimental posture distinguishes EU regulation from traditional command-and-control models. Instead of prescribing substantive outcomes, the EU emphasizes process-oriented obligations. Platforms retain discretion over system design, provided they comply with reporting, auditing, and due diligence requirements. The effectiveness of this approach depends on whether procedural oversight can meaningfully influence automated governance.

Transparency as a Regulatory Strategy:-

Transparency occupies a central role in EU regulation of content moderation. Platforms are required to disclose information about their policies, algorithms, and enforcement practices. These disclosures are intended to enable regulatory oversight, academic research, and public scrutiny. The European Parliament (2020) report highlights transparency as a prerequisite for accountability. Without insight into how automated systems operate, regulators cannot assess their impact on fundamental rights. Transparency obligations thus aim to pierce the opacity that characterizes algorithmic governance. However, transparency is an inherently limited tool. Disclosures often take the form of aggregate statistics, high-level descriptions, or technical documentation that is difficult to interpret. Platforms may comply formally while withholding information that would enable meaningful evaluation. Moreover, transparency does not automatically translate into enforceability. Knowing how a system operates does not necessarily provide mechanisms to correct its failures.

Douek (2022) cautions against equating transparency with accountability. Systems thinking reveals that harms emerge from interactions between components rather than isolated decisions. Transparency focused on individual elements may obscure systemic effects. For example, reporting on takedown rates does not capture how moderation interacts with ranking algorithms, recommendation systems, or economic incentives. Transparency also risks legitimizing automation by framing it as auditable and manageable. When regulators focus on disclosure rather than outcomes, they may inadvertently endorse the underlying system. Transparency becomes a means of documenting harm rather than preventing it.

The Digital Services Act and Systemic Risk:-

The Digital Services Act represents the most comprehensive attempt to regulate platform moderation in the EU. It introduces obligations for very large online platforms to assess and mitigate systemic risks associated with their services, including risks to freedom of expression and civic discourse. Under the DSA, platforms must conduct risk assessments, implement mitigation measures, and submit to independent audits. These requirements reflect an acknowledgment that platform harms are systemic rather than incidental. Automated moderation is treated as a potential source of risk that must be evaluated and managed. Despite this recognition, the DSA remains largely procedural. It does not impose substantive limits on automation or mandate specific design choices. Platforms retain discretion to determine how risks are assessed and mitigated. Regulatory intervention occurs primarily through review and enforcement of compliance obligations. This approach reflects a broader trend toward meta-regulation, in which regulators oversee processes rather than directly controlling outcomes. While meta-regulation offers flexibility and adaptability, it also depends heavily on platform cooperation and regulatory capacity. If regulators lack the expertise or resources to challenge platform assessments, oversight risks becoming superficial. Moreover,

systemic risk assessments are forward-looking and speculative. They rely on platform-provided data and models, which may understate or mischaracterize harms. Without independent access to underlying systems, regulators may struggle to verify claims or enforce meaningful change.

Compliance-Driven Automation:-

EU regulation assumes that automation can be made compatible with fundamental rights through compliance mechanisms. Platforms are encouraged to integrate legal requirements into system design, aligning automation with regulatory expectations. This compliance-driven model treats law as an input into technical systems rather than as an external constraint. Douek (2022) highlights the risks of this approach. When compliance becomes a design parameter, platforms may optimize systems to satisfy formal requirements without addressing substantive harms. For example, platforms may improve documentation and reporting while maintaining aggressive takedown thresholds. Procedural compliance substitutes for normative judgment. Compliance-driven automation also reinforces risk-based enforcement. Platforms seek to demonstrate diligence by erring on the side of removal, reducing regulatory exposure. This tendency aligns with the liability incentives discussed in earlier chapters, perpetuating over-removal and chilling effects. The emphasis on compliance shifts attention away from user experience and rights. Platforms interact primarily with regulators, not affected individuals. Users remain subject to opaque systems with limited remedies, while platforms focus on satisfying audit criteria.

Enforceability and Regulatory Capacity:-

A central question is whether EU regulators possess the capacity to enforce compliance effectively. Transparency and reporting obligations are only meaningful if regulators can interpret, verify, and act on disclosed information. The European Parliament (2020) acknowledges challenges in regulatory capacity, including technical complexity and resource constraints. Automated moderation systems are highly specialized and constantly evolving. Regulators may struggle to keep pace, relying on platform cooperation and external experts. Enforcement mechanisms under the DSA include fines and corrective measures, but their deterrent effect depends on consistent application. Large platforms may treat penalties as costs of doing business, particularly if compliance requirements are ambiguous or negotiable. Enforceability is further complicated by the transnational nature of platforms. While EU law applies within its jurisdiction, platforms operate globally. Automated systems are rarely segmented by region, leading to spillover effects. Regulatory constraints in one jurisdiction may influence global practices, but they may also be diluted by competing obligations elsewhere.

Regulation and the Normalization of Automation:-

A paradox emerges from EU regulatory efforts. By treating automation as inevitable and focusing on procedural oversight, regulation risks normalizing automated moderation as the default mode of speech governance. Rather than questioning whether certain forms of automation are compatible with freedom of expression, regulators seek to manage their risks. This normalization has normative implications. Automation becomes framed as a technical issue subject to optimization rather than a political choice requiring justification. Regulatory discourse shifts from whether automation should govern speech to how it should be audited. Douek (2022) argues that this framing obscures the value judgments embedded in system design. By focusing on transparency and compliance, regulation may legitimize decisions that would be unacceptable if made openly by public authorities. Automation benefits from a form of regulatory laundering, acquiring legitimacy through process rather than substance.

This chapter has examined the European Union's regulatory response to automated content moderation, focusing on transparency, systemic risk, and compliance-based governance. While EU law represents the most ambitious attempt to regulate platform moderation, it remains constrained by structural limitations. Transparency improves visibility but does not guarantee accountability. Systemic risk assessments acknowledge complexity but rely heavily on platform discretion. Compliance-driven automation integrates law into system design without altering underlying incentives. EU regulation functions as a regulatory laboratory, experimenting with process-oriented oversight rather than substantive constraints. While these efforts may mitigate some harms, they do not resolve the legitimacy deficits identified in earlier chapters. Automation remains central to platform governance, insulated by scale, complexity, and economic incentives. The final chapter synthesizes these findings, arguing that automated speech governance faces normative limits that cannot be resolved through transparency or compliance alone. It considers whether automation can ever be reconciled with freedom of expression or whether certain functions must remain beyond its reach.

Conclusion and Discussion:-

Reframing the Problem of Automated Speech Governance:-

This paper set out to examine automated content moderation not as a technical phenomenon, but as a legal and normative problem of governance. Across the preceding chapters, it has shown that platforms now function as private governors of online speech, exercising regulatory power comparable in scope and impact to that of states. Automation has intensified this power by enabling enforcement at unprecedented scale, while simultaneously eroding procedural safeguards, contextual judgment, and accountability. The central claim that emerges from this analysis is not that automated moderation is flawed because it is imperfect. All governance systems are imperfect. Rather, the problem is that automation restructures speech governance in ways that are fundamentally incompatible with the normative foundations of freedom of expression. Efficiency, scalability, and risk management have come to displace legality, fairness, and reasoned judgment as the dominant values shaping moderation systems. This concluding chapter argues that there are normative limits to the automation of speech governance. Certain functions central to the regulation of expression cannot be legitimately delegated to automated systems, regardless of technical sophistication or regulatory oversight. These limits arise from the nature of speech itself, the requirements of legitimacy in governance, and the rule-of-law constraints that apply whenever power is exercised over expression.

Automation and the Absence of Contextual Judgment:-

One of the paper's core findings is that automated moderation lacks the capacity for contextual judgment. Speech is inherently context-dependent. Meaning is shaped by intent, audience, cultural reference, irony, power relations, and political circumstances. Legal systems have long recognized this complexity, which is why restrictions on expression are traditionally assessed through case-by-case analysis rather than mechanical application. Automated systems, by contrast, operate through pattern recognition and probabilistic inference. They do not interpret meaning; they approximate it. As demonstrated in Chapters 2 and 4, this approximation produces inconsistent and biased outcomes, particularly for speech that is ambiguous, critical, or culturally specific. While automation may be effective at identifying clearly unlawful content at scale, it systematically fails at the margins, where lawful and socially valuable expression is most vulnerable.

The normative problem is not simply that machines make mistakes. It is that automation replaces interpretive judgment with statistical correlation. Decisions affecting expression are no longer grounded in reasons that can be articulated, challenged, or reviewed, but in outputs that are opaque by design. This shift undermines the justificatory logic of speech regulation, which depends on the ability to explain why a particular restriction is warranted. From a normative perspective, governance over speech requires more than accuracy. It requires judgment that is sensitive to context, proportion, and social meaning. Automation, as currently deployed, cannot meet this requirement. This limitation is structural rather than temporary, and it marks a boundary beyond which automation cannot legitimately extend.

Legitimacy Depends on Fairness, Not Scale:-

A recurring justification for automated moderation is scale. Platforms process vast volumes of content, making individualized assessment appear infeasible. Automation is presented as a necessary response to the realities of digital communication. This paper has shown that while scale explains the turn to automation, it cannot justify the abandonment of fairness. Legitimacy in governance does not derive from efficiency alone. In both public law and human rights traditions, legitimacy depends on the fairness of procedures, the transparency of decision-making, and the availability of effective remedies. As Chapters 3 and 5 demonstrated, automated moderation systems systematically fail to provide these safeguards. Notice is opaque, explanations are absent, and appeals are dysfunctional or illusory.

Are's analysis of appeals failures underscores a crucial point: procedural fairness is not an optional enhancement but a constitutive element of just governance. A system that produces correct outcomes some of the time but denies individuals the ability to understand or challenge decisions lacks legitimacy, regardless of its efficiency. Scale does not negate this requirement. If anything, it heightens it. When governance affects billions of users and shapes global public discourse, the need for fairness becomes more acute, not less. The argument that due process must be sacrificed for scale reflects a prioritization of institutional convenience over individual rights. Normatively, this trade-off is indefensible.

The Illusion of Neutral and Inevitable Automation:-

Another theme running through this paper is the illusion of neutrality that surrounds automated moderation. Automation is often framed as objective, consistent, and inevitable. This framing obscures the normative choices embedded in system design, including decisions about what to prioritize, what to risk, and whose speech to protect. As Klonick's work makes clear, platforms borrow the language and structure of law without adopting its accountability mechanisms. Automation intensifies this dynamic by presenting governance decisions as technical outputs rather than value-laden judgments. When moderation outcomes are attributed to algorithms, responsibility is diffused and critique is deflected. This paper rejects the notion that automation is inevitable. Automation is a choice, shaped by economic incentives, liability regimes, and regulatory design. Treating it as unavoidable forecloses normative debate and legitimizes outcomes that would be unacceptable if openly acknowledged as discretionary decisions. Recognizing automation as a contingent governance choice is essential to establishing its limits. If automation is not inevitable, it can be constrained, redesigned, or excluded from certain domains. Normative analysis must therefore focus not on how to perfect automated systems, but on where they should not be used at all.

Subordination of Automation to Law:-

The final normative claim advanced by this paper is that automation must remain subordinate to law. This does not mean that automated tools have no role in content moderation. It means that they cannot be allowed to define the boundaries of permissible speech independently of legal and human rights standards. Subordination requires more than transparency or compliance reporting. As Chapter 6 demonstrated, regulatory frameworks that focus on procedural oversight risk legitimizing automation without altering its underlying logic. When law is treated as an input into system design rather than an external constraint, automation absorbs legal norms without being meaningfully constrained by them. To remain subordinate to law, automated moderation must be limited in scope and function. Decisions that require contextual interpretation, balancing of competing rights, or assessment of social meaning must involve human judgment accountable to legal standards. Automated systems may assist in identifying potential issues, but they cannot serve as final arbiters of expression.

This normative boundary reflects a core rule-of-law principle: coercive power over expression must be exercised through processes that are intelligible, contestable, and accountable. Automation, as currently implemented, does not satisfy these conditions. Allowing it to operate autonomously in speech governance undermines the very values that freedom of expression is meant to protect. Taken together, the chapters of this paper advance a unified argument. Platforms have become private governors of online speech. Automation has transformed moderation into a system of large-scale regulation driven by risk and efficiency. Procedural safeguards have eroded, substantive arbitrariness has become systemic, and human rights constraints struggle to operate in a privatized, automated environment. Regulatory responses, while ambitious, often manage rather than challenge these dynamics. The contribution of this paper lies in articulating the normative limits of automation in speech governance. Rather than proposing incremental reforms or technical fixes, it draws a conceptual boundary around what automation can legitimately do. It argues that certain aspects of speech regulation; contextual judgment, reason-giving, and individualized fairness; are incompatible with full automation.

This boundary does not deny the challenges of governing speech at scale. It insists, however, that scale cannot serve as a justification for abandoning the normative foundations of expressive freedom. Where automation cannot meet those foundations, it must yield to law. The rise of automated content moderation represents a profound transformation in the governance of expression. What was once the domain of courts and public authorities is increasingly managed by private systems optimized for speed and risk avoidance. This paper has shown that such systems are not merely imperfect but normatively misaligned with the values they are tasked with protecting. The question, therefore, is not whether automation can be improved, but whether it should govern speech at all in certain contexts. Freedom of expression has always required friction: deliberation, disagreement, and justification. Automation promises frictionless governance. The normative lesson of this paper is that friction is not a flaw in speech governance, but a feature essential to its legitimacy. Automation may assist governance, but it cannot replace judgment. Where speech is concerned, that limit matters.