Monetizing Uncertainty: Institutional Foreseeability and the Political Economy of Platform Fraud

Flyxion

January 17, 2026

Abstract

Large online platforms frequently characterize fraud, impersonation, and scam activity as unavoidable externalities of scale. This framing presumes partial ignorance: that harmful behavior emerges faster than it can be detected or meaningfully constrained. Drawing on internal documents reviewed by Reuters in late 2025, this paper challenges that assumption in the case of Meta Platforms. The documents reveal that Meta internally models scam exposure at the scale of tens of billions of daily impressions and projects that a substantial share of its advertising revenue derives from scams and other prohibited goods. These estimates demonstrate not only awareness but quantitative foreseeability.

This paper introduces the concept of *monetized uncertainty* to describe governance regimes in which probabilistic harm is not eliminated but priced, bounded, and rendered economically productive. It argues that Meta's enforcement architecture—characterized by high fraud-confidence thresholds, penalty pricing for suspected offenders, disposable identity systems, and revenue guardrails— constitutes revenue-calibrated enforcement rather than harm prevention. Under such conditions, fraud persists as a stable equilibrium because anticipated regulatory penalties remain smaller than the revenue generated by illicit activity.

Situating this analysis within platform economics, corporate compliance theory, content moderation scholarship, and fraud victimology, the paper demonstrates that monetized uncertainty is not an idiosyncratic failure but a predictable outcome of advertiser-funded platforms operating at scale. It further argues that user-level controls and incremental moderation improvements are structurally insufficient when identity is unbound from history and enforcement resets rather than accumulates.

The paper concludes by proposing a constraint-first framework for platform governance grounded in history-bound identity, irreversible enforcement, and the decoupling of revenue from probabilistic harm. Under such a framework, uncertainty is resolved in favor of safety rather than profit, rendering large-scale fraud economically irrational rather than strategically optimal.

1 Introduction

Large online platforms routinely characterize fraud, impersonation, and scam activity as unavoidable externalities of scale. This characterization frames abuse as an emergent property of complex systems rather than as the predictable outcome of institutional design choices. Within this framing, harm appears as something regrettable yet fundamentally ungovernable, arising faster than it can be detected, classified, or prevented. Such claims depend on an assumption of partial ignorance: that platforms lack sufficient information to distinguish legitimate participation from abuse in advance, and must therefore rely on reactive enforcement.

Recent disclosures regarding Facebook and its parent company, Meta Platforms, challenge this assumption directly. Internal documents reviewed by Reuters in late 2025 demonstrate not only awareness of large-scale fraudulent activity across Meta's platforms, but detailed quantitative modeling of its volume, persistence, and revenue contribution [1]. According to these documents, Meta internally estimated that its users are exposed to tens of billions of scam attempts each day, including approximately fifteen billion paid scam advertisements. The company further projected that roughly ten percent of its annual advertising revenue derives from scams and other prohibited goods. These estimates were produced internally, across finance, safety, engineering, and lobbying divisions, and were incorporated into revenue forecasts and risk assessments.

The existence of such modeling fundamentally alters the governance question. When harmful activity is internally quantified, forecast, and incorporated into business planning, its persistence can no longer be attributed to ignorance, technical incapacity, or adversarial evasion alone. Instead, it reflects a choice about how much harm to tolerate under conditions of uncertainty, and how to balance that harm against revenue and regulatory exposure. This essay argues that Meta's approach to scam advertising exemplifies a broader institutional pattern that can be described as *monetized uncertainty*: a governance regime in which probabilistic harm is not eliminated, but priced, managed, and rendered economically productive.

This argument situates Meta's conduct within the political economy of advertiser-funded platforms, drawing on scholarship in platform economics, content moderation, corporate compliance, and fraud victimology. It contends that the persistence of large-scale fraud on social platforms is not an incidental failure of moderation, but a predictable equilibrium produced by incentive structures that reward partial enforcement while penalizing decisive intervention. In such systems, moderation functions less as a mechanism of protection than as a form of reputational risk management calibrated to preserve revenue streams.

2 Formal Framework: Monetized Foreseeability

To analyze this governance pattern with precision, it is necessary to formalize the relationship between internal knowledge, revenue generation, and enforcement decisions. The concept of monetized uncertainty captures a specific institutional condition in which harmful activity is both foreseeable and profitable, and in which enforcement is optimized not to eliminate harm, but to bound its external consequences.

Let H denote a class of harmful activity, such as scam advertising, and let R_H denote the revenue derived from H. A platform exhibits monetized foreseeability with respect to H when it satisfies three conditions. First, the platform maintains internal models that estimate the frequency, scale, or exposure of H with sufficient granularity to inform business decisions. Second, the platform quantifies R_H and incorporates it into financial projections, risk assessments, or strategic planning. Third, enforcement policies are calibrated such that R_H remains positive while regulatory, legal, and reputational costs are constrained within acceptable bounds.

Under these conditions, enforcement decisions are not driven by the objective of minimizing H, but by an optimization problem in which H is treated as a probabilistic revenue source subject to constraint. Formally, enforcement thresholds are selected to maximize expected profit net of reputational and regulatory costs, rather than to minimize harm directly. This framework aligns with analyses of compliance behavior in other industries, where firms facing diffuse harms and bounded penalties routinely price misconduct into operational strategy rather than eliminate it [23, 25].

In the context of platform governance, this optimization is facilitated by the probabilistic nature of content classification. Fraud detection systems rarely produce binary determinations; instead, they assign likelihoods. The critical governance decision, therefore, lies not in whether fraud can be detected, but in where enforcement thresholds are set. Meta's internal documents indicate that advertisers are removed only when automated systems assign a fraud likelihood exceeding ninety-five percent [1]. Below that threshold, advertisers assessed as likely but not sufficiently certain to be fraudulent are permitted to continue operating, often subject to higher advertising costs rather than exclusion.

This threshold-based regime transforms uncertainty itself into a revenue-generating asset. By allowing advertisers in the ambiguous zone between suspicion and certainty to continue purchasing ads, the platform monetizes classification uncertainty while maintaining plausible deniability. Harm is not denied; it is statistically bounded. Enforcement becomes a matter of tuning parameters rather than exercising judgment, and responsibility is displaced from institutional design to algorithmic output.

3 Identity Without History

The effectiveness of this regime depends critically on the treatment of identity within the platform. Meta's advertising and page infrastructure permits rapid identity creation, minimal historical binding, and low-cost abandonment. Accounts function as interchangeable vessels rather than as persistent actors. When enforcement occurs, it removes particular accounts or pages, not the underlying entities controlling them. New identities can be created with negligible cost, allowing abusive actors to reenter the system almost immediately.

This absence of historical accumulation undermines deterrence. In systems where identity carries no durable consequence, punishment loses its force. Enforcement actions become episodic interruptions rather than meaningful constraints, and the expected cost of removal is easily outweighed by short-term gains. Platform economics scholarship has long noted that two-sided markets amplify such asymmetries, as legitimate advertisers invest in long-term reputation while fraudulent actors optimize for rapid extraction before detection [5, 6, 7]. When identity is disposable, this asymmetry becomes structurally stable rather than self-correcting.

Meta's internal modeling implicitly acknowledges this dynamic. Documents reviewed by Reuters indicate that some high-spending advertisers accrued hundreds of enforcement strikes without being removed, and that enforcement teams were explicitly constrained by revenue guardrails limiting the financial impact of moderation actions [1]. In such a context, identity churn is not a failure mode but an assumed operating condition. The system is designed to tolerate recidivism so long as it remains economically productive.

4 From Moderation to Risk Management

The combination of probabilistic enforcement thresholds and disposable identity produces a distinctive form of governance in which moderation operates primarily as reputational risk management. Harm to users is treated as a degradation of experience, while harm to advertisers, public figures, or regulators is treated as a business risk requiring intervention. This asymmetry is reflected in enforcement priorities that focus disproportionately on brand impersonation and high-profile abuse, while user-reported scams are frequently ignored or dismissed [1, 2].

Such patterns are consistent with broader analyses of content moderation in advertiser-funded platforms. Gillespie has shown that moderation systems are shaped less by normative commitments than by economic dependencies, producing enforcement practices that protect revenue concentration rather than diffuse user welfare [9]. When user harm is widespread but individually small, and advertiser harm is concentrated and reputationally salient, platforms predictably allocate enforcement resources toward the latter.

Within this framework, claims that fraud represents an unavoidable externality lose credibility. The persistence of harm is not merely foreseeable; it is anticipated, modeled, and bounded. The relevant question is therefore not whether Meta can do better, but why it chooses not to, and what incentive structures render that choice rational.

5 Revenue-Calibrated Enforcement

Meta's internal documents reveal that enforcement against fraudulent advertisers is governed by explicit probabilistic thresholds rather than categorical prohibitions. Advertisers are removed only when automated systems assign a likelihood of fraud exceeding ninety-five percent. Below this threshold, advertisers assessed as likely—but not sufficiently certain—to be engaged in fraud are permitted to continue operating, often under modified economic terms such as increased advertising costs or restricted delivery [1]. This design choice is frequently defended as a necessary safeguard against false positives, preserving access for legitimate advertisers whose content may resemble fraudulent material.

However, the presence of alternative enforcement modalities undermines this justification. Platforms possess a wide range of graduated responses that could reduce harm without permanent exclusion, including manual review requirements, escrowed payments, delayed campaign launches, or escalating verification burdens. Meta's choice to implement penalty pricing rather than exclusion indicates that the governing objective is not harm minimization but revenue preservation under uncertainty.

From an economic perspective, this approach closely resembles practices observed in financial markets prior to major regulatory reforms, where institutions priced known risks into profit calculations rather than eliminating them [25]. In such environments, misconduct persists not because it is undetectable, but because its expected returns exceed its expected costs. Meta's internal analyses reportedly show that increased prices paid by suspected fraudsters offset revenue losses from removed ads, resulting in a net preservation of income even as enforcement intensifies marginally [1].

This revenue-calibrated enforcement regime transforms moderation into an optimization problem. Enforcement thresholds are tuned to balance incremental reductions in harm against marginal revenue loss and anticipated reputational or regulatory consequences. The result is a stable equilibrium in which some level of fraud is not merely tolerated, but economically integrated into platform operations.

6 The Economic Stabilization of Harm

The persistence of large-scale fraud on Meta's platforms cannot be understood solely through enforcement policy; it must also be examined through the lens of regulatory asymmetry. Internal documents reviewed by Reuters indicate that Meta anticipates regulatory fines of up to one billion dollars related to scam advertising [1]. At the same time, the company estimates annual revenue from high-risk scam advertisements alone at several multiples of that figure.

This disparity produces a predictable outcome. When anticipated penalties remain smaller than the revenue derived from harmful activity, enforcement efforts are rationally constrained to avoid significant revenue disruption. Regulatory sanctions become a cost of doing business rather than a deterrent. Similar dynamics have been documented across industries where compliance failures are penalized episodically and at levels insufficient to alter core business incentives [23, 24].

Historical precedent reinforces this interpretation. In 2019, the Federal Trade Commission imposed a five-billion-dollar penalty on Facebook for privacy violations [19]. While substantial in absolute terms, the fine represented a small fraction of Meta's annual revenue and did not fundamentally alter its advertising-driven business model. Subsequent regulatory actions have followed a similar pattern, imposing fines that attract public attention without exceeding the financial gains produced by the underlying practices.

Under these conditions, harm becomes economically stabilized. Enforcement does not converge toward elimination but toward an equilibrium determined by the ratio of illicit revenue to expected penalties. As long as revenue from fraud exceeds the cost of regulatory intervention, harmful activity persists at a predictable steady state. The system does not fail to correct itself; it succeeds in optimizing within misaligned constraints.

7 User-Level Controls and the Illusion of Agency

Platforms frequently emphasize user-level tools—blocking, reporting, and content filtering—as evidence of their commitment to safety. These mechanisms presume that abuse is relatively rare, that individual users can meaningfully reduce exposure through vigilance, and that collective reporting will trigger timely enforcement. Meta's internal estimates invalidate these assumptions.

According to internal documents, Meta's platforms expose users to tens of billions of scam attempts daily, including both paid advertisements and so-called "organic" scams that operate through direct messages, marketplace listings, and group activity [1]. At such scales, individual reporting becomes statistically irrelevant. Even if users correctly identify and report fraudulent content, enforcement systems are overwhelmed by volume and constrained by revenue considerations.

Moreover, research on fraud victimology demonstrates that narratives emphasizing individual responsibility systematically underestimate the sophistication of modern scams and the asymmetric information environments in which they operate [12, 14]. Fraudulent campaigns are often adaptive, personalized, and socially engineered to exploit trust, authority, and urgency. When platforms possess comprehensive behavioral data and detection infrastructure, shifting responsibility onto users constitutes a misallocation of accountability rather than a realistic governance strategy.

Meta's own internal data indicate that the vast majority of valid user reports of scams are ignored or incorrectly rejected [1]. Under such conditions, user-level controls function primarily as symbolic gestures. They provide a sense of agency without materially reducing harm, reinforcing the perception that responsibility lies with users rather than with institutional design.

8 The Myth of Technical Infeasibility

A common defense of platform tolerance for fraud invokes technical infeasibility. According to this view, sophisticated adversaries evolve faster than detection systems, rendering comprehensive enforcement impossible. This argument collapses under scrutiny when examined alongside Meta's internal modeling.

If a platform can estimate scam exposure at the level of billions of impressions per day, categorize revenue contributions with percentage-level precision, and differentiate between classes of fraud for business planning purposes, it possesses sufficient detection capability to enforce more aggressively. The limitation is not epistemic but economic. Enforcement thresholds are set not at the limits of technical possibility, but at points that preserve revenue while containing reputational risk.

This conclusion aligns with prior disclosures regarding Meta's internal decision-making. Frances Haugen's testimony before the U.S. Senate documented systematic deprioritization of user safety in favor of engagement and revenue metrics, even when internal research demonstrated foreseeable harm [2]. The pattern revealed by the Reuters documents represents a continuation of this logic, applied specifically to fraudulent advertising.

Technical infeasibility thus serves as a rhetorical shield rather than an empirical explanation. It reframes economic choices as engineering constraints, obscuring the role of incentive alignment in shaping enforcement outcomes.

9 Moderation as Compliance Theater

Taken together, these dynamics produce what corporate governance scholars have described as compliance theater: the appearance of robust enforcement without substantive commitment to elimination [23]. Platforms publicize takedown numbers, emphasize investments in artificial intelligence, and highlight reductions in reported incidents, while maintaining enforcement architectures that permit large-scale harm to persist.

Meta's internal celebration of initiatives such as "Scammiest Scammer" reports and penalty bidding schemes illustrates this phenomenon [1]. Such measures generate internal acknowledgment of the problem and external narratives of action, yet often fail to result in durable exclusion of high-impact offenders. Accounts identified as major sources of abuse may remain active for months, continuing to generate revenue until external scrutiny forces intervention.

Compliance theater is particularly effective in regulatory environments characterized by fragmented oversight and delayed enforcement. Platforms can demonstrate good-faith effort through process metrics while avoiding structural changes that would disrupt revenue. Over time, this produces a governance regime in which harm reduction is perpetually promised but rarely realized.

10 Reframing Responsibility

The persistence of fraud under conditions of internal foreseeability necessitates a reframing of responsibility. When platforms possess the capacity to model harm, estimate its revenue contribution, and optimize enforcement accordingly, responsibility for resulting damage cannot be plausibly assigned to individual users or isolated bad actors. It resides instead in institutional design choices that render harm profitable.

This reframing aligns with emerging legal and ethical scholarship challenging the scope of intermediary immunity when platforms materially contribute to harm through monetization and optimization [17, 10]. It also resonates with broader critiques of surveillance capitalism, which emphasize the extraction of value from behavioral uncertainty as a core business strategy [8]. In such systems, uncertainty is not an obstacle to governance but a resource to be exploited.

The following sections extend this analysis beyond Meta, situating its practices within a comparative platform landscape and examining the legal and regulatory frameworks that enable monetized uncertainty to persist.

11 Comparative Platform Analysis

Although this essay focuses on Meta due to the unusual availability of internal documents, the governance pattern it illustrates is not unique. Rather, it reflects structural features common to advertiser-funded platforms operating at scale. Comparative analysis across major platforms reveals convergent strategies for managing probabilistic harm under conditions of revenue dependence.

Search-based platforms such as Google have faced longstanding criticism for hosting fraudulent advertisements, including phishing schemes, counterfeit goods, and deceptive financial products. Public transparency reports indicate that billions of advertisements are removed annually, yet enforcement remains largely reactive, triggered by complaints or external scrutiny rather than preventative exclusion. Sponsored search auctions exhibit similar dynamics to Meta's advertising systems, with suspected bad actors often subjected to increased costs or quality score penalties rather than outright exclusion. As in Meta's case, uncertainty is monetized through differential pricing rather than resolved through structural deterrence.

Video-centric and recommendation-driven platforms exhibit parallel vulnerabilities. TikTok's rapid growth has coincided with an increase in scam advertising and deceptive influencer marketing, particularly targeting younger users. Algorithmic recommendation systems amplify engagement-maximizing content regardless of legitimacy, creating fertile ground for fraudulent campaigns that exploit novelty, urgency, and social proof. Despite differences in interface and demographic reach, these platforms share a reliance on scale, automation, and probabilistic enforcement that produces similar outcomes.

What distinguishes Meta is therefore not the existence of monetized uncertainty, but the degree to which it has been internally documented and quantified. The convergence observed across platforms suggests that the persistence of fraud is a predictable consequence of advertiser-funded business models rather than the idiosyncratic failure of a single firm. Where revenue depends overwhelmingly on advertising volume, enforcement incentives align toward partial tolerance rather than elimination.

12 Structural Commonalities Across Platforms

Across diverse platforms, several structural features recur. Advertising revenue dominates platform income, creating pressure to maximize participation while minimizing friction. Enforcement systems operate probabilistically, producing confidence scores rather than categorical judgments. Identity systems prioritize ease of entry and scale over historical continuity. Regulatory oversight is fragmented, delayed, and often jurisdiction-specific, allowing firms to arbitrage enforcement intensity across markets.

These features interact to produce a stable governance equilibrium. Harmful activity persists not because it is invisible, but because its elimination would require interventions that conflict with revenue optimization. In this equilibrium, uncertainty itself becomes valuable. Probabilistic detection enables platforms to claim diligence while retaining flexibility, and ambiguous cases generate revenue that would be foregone under stricter regimes.

This convergence supports the broader claim that monetized uncertainty is an emergent property of platform political economy rather than a contingent managerial failure. Addressing it therefore requires interventions that alter incentive structures rather than incremental improvements in detection accuracy.

13 Legal Architecture and Foreseeability

The persistence of monetized uncertainty is reinforced by existing legal frameworks, particularly in the United States. Section 230 of the Communications Decency Act provides platforms with broad immunity from liability for third-party content, even when they possess knowledge of harmful activity. Courts have generally interpreted this immunity expansively, shielding platforms from responsibility so long as they do not materially contribute to the creation of unlawful content.

This doctrine becomes strained, however, when platforms actively monetize harm they can foresee and quantify. Legal scholars have argued that immunity should not extend to situations in which platforms knowingly profit from harmful conduct and calibrate enforcement to preserve that profit. When internal documents demonstrate that harm is modeled, priced, and incorporated into strategic planning, the distinction between passive intermediation and active participation becomes increasingly difficult to sustain.

European regulatory approaches have begun to reflect this tension. The Digital Services Act imposes affirmative obligations on large platforms to assess and mitigate systemic risks, including fraud. Unlike U.S. frameworks that emphasize liability shields, the DSA requires platforms to demonstrate proportionality between identified risks and enforcement measures. Internal modeling of scam exposure and revenue contribution would likely constitute evidence of foreseeable systemic risk under this regime.

Early enforcement actions suggest growing skepticism toward moderation strategies that preserve revenue while offering symbolic compliance. By shifting the regulatory focus from individual content decisions to systemic risk management, European regulators have begun to challenge the economic logic underpinning monetized uncertainty. Whether these efforts will meaningfully alter platform incentives remains an open question.

14 Foreseeability and Corporate Governance

From a corporate governance perspective, Meta's internal documentation aligns with well-established patterns in organizational behavior. Firms operating under diffuse harm and bounded penalties frequently adopt compliance strategies that minimize legal exposure without eliminating underlying misconduct. This behavior is not anomalous; it is rational within incentive structures that reward short-term profitability and treat fines as manageable costs.

What distinguishes Meta's case is the scale and granularity of internal knowledge. The ability to estimate scam exposure in the tens of billions of daily impressions, project revenue contributions with precision, and anticipate regulatory penalties demonstrates a level of foresight that exceeds the threshold traditionally associated with plausible deniability. Under such conditions, continued monetization of harm reflects an institutional choice rather than an informational deficit.

This reframing has significant implications for accountability. It suggests that debates over moderation efficacy miss the central issue, which is not whether platforms can detect fraud more accurately, but whether they are willing to accept the revenue consequences of doing so. As long as enforcement remains subordinate to revenue optimization, improvements in detection will merely refine the pricing of uncertainty rather than eliminate harm.

15 Toward a Different Analytical Lens

The preceding analysis indicates that monetized uncertainty cannot be addressed solely through technical fixes, expanded moderation teams, or incremental policy changes. It arises from deeper assumptions about identity, reversibility, and constraint that shape platform architecture. To move beyond compliance theater, it is necessary to reconsider these foundational assumptions.

In particular, platforms treat identity as transient and actions as reversible. Accounts can be created, removed, and recreated with minimal consequence. Enforcement resets rather than accumulates. Under such conditions, harm diffuses rather than dissipates, and the system's entropy increases over time.

The following section introduces an alternative interpretive lens grounded in history-bound identity and irreversible action. This lens provides a framework for understanding why monetized uncertainty persists and how platform governance might be restructured to render harm economically irrational rather than strategically optimal.

16 History, Irreversibility, and the Accumulation of Constraint

The persistence of monetized uncertainty can be understood more clearly when examined through the lens of history and irreversibility. Contemporary platforms are designed as if actions were fundamentally reversible and identities were stateless. Accounts are created, suspended, and recreated; advertisers are removed and reintroduced; enforcement resets rather than compounds. In such systems, harm does not meaningfully decay. It disperses, recombines, and reappears under new identifiers.

By contrast, systems that successfully regulate harmful behavior treat history as an irreducible substrate. Actions alter future possibilities, and repeated violations increase constraint rather than reset it. Identity, under these conditions, is not merely a label but a trajectory through time, shaped by irreversible events. Trust emerges not as a subjective judgment but as an informational property of accumulated history.

Meta's governance architecture largely lacks this property. Enforcement actions do not meaningfully alter the future state space available to abusive actors. Removal is not an irreversible event but a transient interruption. As a result, deterrence fails not because punishment is insufficiently severe, but because it is insufficiently cumulative. Without historical binding, the system cannot learn in a way that constrains future behavior.

This absence of irreversibility explains why improvements in detection accuracy do not translate into proportional reductions in harm. Better classifiers simply sharpen the probabilistic boundary at which monetization occurs. They do not alter the underlying dynamics that allow harm to regenerate.

17 Entropy, Coherence, and Institutional Design

The concept of entropy provides a useful metaphor for understanding platform governance failure. In high-entropy systems, distinctions blur, identities fragment, and signals degrade into noise. When harmful activity can be continuously reintroduced at low cost, the system expends increasing effort merely to maintain its current state, let alone improve it.

Meta's moderation apparatus exhibits precisely this pattern. Vast resources are devoted to identifying and removing individual instances of abuse, yet overall exposure remains high. The system does not converge toward lower harm; it oscillates around a stable equilibrium determined by economic constraints. Entropy is managed locally but allowed to accumulate globally.

Coherence, by contrast, requires constraint. Systems that reduce entropy do so by limiting the space of permissible transitions. In social and institutional contexts, this means binding identity to history, making certain actions costly or irreversible, and ensuring that repeated violations progressively restrict future participation. Without such constraints, governance becomes reactive and Sisyphean.

From this perspective, monetized uncertainty is not merely an ethical failure but a thermodynamic one. The platform expends energy moderating content without altering the structural conditions that generate it. Harm persists because the system's design allows it to do so.

18 Constraint-First Design Principles

If monetized uncertainty arises from the absence of binding constraints, then meaningful reform must begin at the level of architectural design rather than policy overlay. Constraint-first governance treats limits not as after-the-fact interventions but as foundational elements that shape behavior from the outset.

One such constraint concerns identity. Platforms could require that advertisers and high-impact accounts accrue irreversible history, such that repeated enforcement actions increase the cost of continued participation. Identity would cease to be a disposable container and become a cumulative record. Re-entry would no longer reset the system but tighten it.

A second constraint concerns monetization pathways. Enforcement and revenue functions must be structurally decoupled. Systems that profit from probabilistic harm cannot credibly govern it. Where uncertainty exists, default responses should reduce reach or require verification, not extract higher rents. Ambiguity should be treated as a reason for caution rather than an opportunity for pricing.

A third constraint concerns time. Rapid re-entry enables abuse to scale faster than enforcement can respond. Introducing temporal friction—delays, cooling-off periods, or graduated reactivation thresholds—would alter the cost-benefit calculus of fraudulent campaigns. Harm that depends on speed and novelty would lose its advantage.

These constraints do not require perfect detection. They require that uncertainty be resolved in favor of safety rather than profit. By altering the economic landscape in which actors operate, constraint-first design shifts incentives without relying on constant intervention.

19 Institutional Implications

The adoption of constraint-first principles would have implications beyond Meta. It would challenge the prevailing assumption that scale and openness must come at the expense of accountability. It would also undermine the business logic that treats regulatory penalties as external costs rather than as signals of structural misalignment.

From a governance standpoint, such reforms align with emerging regulatory approaches that emphasize systemic risk over individual violations. Rather than mandating specific moderation outcomes, regulators could require platforms to demonstrate that their architectures impose cumulative constraints on harmful behavior and that enforcement decisions are not subordinated to revenue preservation.

Crucially, this approach reframes the debate. The question is no longer whether platforms can eliminate all fraud, but whether they have designed systems in which fraud remains economically rational. Where harm persists because it is profitable, responsibility lies with institutional design rather than with users or adversaries.

20 Conclusion: Beyond Monetized Uncertainty

Meta's internal documents reveal a system that does not merely tolerate uncertainty, but monetizes it. Fraud is not an unforeseen side effect; it is a modeled variable, priced into revenue forecasts and bounded by enforcement thresholds. Under such conditions, appeals to technical limitation or user responsibility are insufficient.

This essay has argued that monetized uncertainty arises from a combination of probabilistic enforcement, disposable identity, and misaligned incentives. Addressing it requires a shift from reactive moderation to constraint-first design, in which history binds identity, irreversibility accumulates consequence, and uncertainty is resolved in favor of safety rather than profit.

Absent such changes, platform governance will continue to oscillate between scandal and reassurance, enforcement and relapse. Harm will remain visible yet unresolved, foreseeable yet tolerated. The challenge, therefore, is not to improve moderation at the margins, but to redesign the systems that make monetized uncertainty the rational equilibrium.

References

[1] J. Horwitz. Meta is earning a fortune on a deluge of fraudulent ads, documents show. *Reuters Special Report*, November 6, 2025 (updated December 28, 2025).

[2] F. Haugen. Testimony before the Senate Committee on Commerce, Science, and Transportation. U.S. Senate, October 5, 2021.

[3] J. Horwitz and D. Seetharaman. Facebook knows Instagram is toxic for teen girls, company documents show. *The Wall Street Journal*, September 14, 2021.

[4] J. Angwin and T. Parris Jr. Facebook's secret rulebook for policing global political speech. *ProPublica*, December 28, 2021.

[5] J.-C. Rochet and J. Tirole. Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029, 2003.

[6] D. S. Evans. Some empirical aspects of multi-sided platform industries. *Review of Network Economics*, 2(3):191–209, 2003.

[7] A. Hagiu and J. Wright. Multi-sided platforms. *International Journal of Industrial Organization*, 43:162–174, 2015.

[8] S. Zuboff. *The Age of Surveillance Capitalism.* PublicAffairs, New York, 2019.

[9] T. Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, New Haven, 2018.

[10] K. Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6):1598–1670, 2018.

[11] S. T. Roberts. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press, New Haven, 2019.

[12] M. Button, C. M. Lewis, and J. Tapley. Not a victimless crime: The impact of fraud on individual victims and their families. *Security Journal*, 27(1):36–54, 2014.

[13] M. T. Whitty and T. Buchanan. The online romance scam: A serious cybercrime. *Cyberpsychology, Behavior, and Social Networking*, 15(3):181–183, 2012.

[14] C. Cross. No laughing matter: Blaming the victim of online fraud. *International Review of Victimology*, 21(2):187–204, 2015.

[15] Federal Trade Commission. Consumer Sentinel Network Data Book 2022. FTC Report, February 2023.

[16] 47 U.S.C. § 230. Protection for private blocking and screening of offensive material. Communications Decency Act of 1996.

[17] D. K. Citron and B. A. Wittes. The Internet will not break: Denying bad Samaritans § 230 immunity. *Fordham Law Review*, 86:401–423, 2014.

[18] J. Kosseff. *The Twenty-Six Words That Created the Internet.* Cornell University Press, Ithaca, 2019.

[19] Federal Trade Commission. FTC imposes $5 billion penalty and sweeping new privacy restrictions on Facebook. Press release, July 24, 2019.

[20] Federal Trade Commission. FTC proposes blanket prohibition preventing Facebook from monetizing youth data. Press release, May 3, 2023.

[21] European Parliament and Council. Regulation (EU) 2022/2065 on a single market for digital services (Digital Services Act). *Official Journal of the European Union*, L 277:1–102, 2022.

[22] Irish Data Protection Commission. Decision in the matter of Facebook Ireland Limited. Reference No. IN-18-12-2, December 2022.

[23] D. C. Langevoort. Cultures of compliance. *American Criminal Law Review*, 54:933–977, 2017.

[24] J. Armour, J. N. Gordon, and G. Min. Taking compliance seriously. *Yale Journal on Regulation*, 37:1–88, 2020.

[25] F. Partnoy. *Infectious Greed: How Deceit and Risk Corrupted the Financial Markets.* PublicAffairs, New York, 2009.