Identity Collapse and the Platforming of Fraud: A Formal and Empirical Critique of Trust Degradation in Facebook-Scale Systems

Flyxion

Abstract

This essay argues that persistent fraud, impersonation, and institutional trust degradation on Facebook-scale social platforms are not adequately explained as incidental "content moderation failures." Instead, they arise predictably from architectural choices that relax identity constraints, detach symbols from histories, and optimize distribution for engagement rather than verification. The manuscript strengthens this claim along three axes: a formal systems interpretation of identity and reference, an information-theoretic and thermodynamic analogy for trust decay, and an empirical program grounded in platform enforcement reporting and investigative documentation. The central thesis is constraint-first: without namespace integrity and event-historical accountability, optimization for engagement produces an environment in which deception becomes a stable equilibrium (Akerlof 1970; Lessig 1999; Wu 2016). The essay concludes with measurable audit metrics, an implementation roadmap for event-historical architectures, and draftable regulatory requirements consistent with existing legal frameworks such as the EU Digital Services Act and related transparency obligations (Regulation (EU) 2022/2065; Regulation (EU) 2016/679).

1 Introduction

Social trust online is a technical and institutional achievement. It depends on stable reference, persistent identity, and mechanisms that allow communities to detect and contain deception at scale. When those conditions are absent, fraud is not merely possible; it becomes adaptive. This essay examines Facebook-scale architecture through that lens and argues that a large portion of the harm users experience—impersonation, scam recurrence, metric laundering, and attention diversion—is best understood as a predictable consequence of weakened identity constraints coupled to engagement-optimized ranking and monetization.

The analysis is not a claim that fraud is unique to any one platform, nor that scale alone implies malice. It is a claim about structural sufficiency: if a platform permits non-unique representational identity, allows identity to shed history at low cost, and rewards distribution by engagement, then impersonation-based fraud and trust degradation will occur at high prevalence even under substantial moderation effort. Economic theory already provides an archetype for this dynamic: when quality signals are unreliable, low-quality goods and deception can drive out honest participation (Akerlof

1970). The attention economy supplies a second archetype: when business models depend on time-on-platform and impression volume, systems tend to optimize extraction rather than epistemic quality (Wu 2016; Zuboff 2019).

This manuscript also treats the problem as empirically testable rather than purely rhetorical. Meta publishes recurring integrity and enforcement reporting that, while incomplete, provides quantities relevant to identity integrity and adversarial pressure. For example, Meta has repeatedly estimated fake-account prevalence on Facebook at roughly a few percent of monthly active users and reports actioning large volumes of fake accounts in quarterly windows, implying an ongoing, industrial-scale adversarial environment rather than sporadic abuse (Meta 2018; Meta Transparency Centre 2025). In parallel, investigative reporting has documented allegations, supported by internal documents, that scam advertising is a persistent revenue and enforcement challenge for Meta platforms, and that enforcement choices are shaped by business incentives as well as technical uncertainty (Reuters 2025a; Reuters 2025b; Reuters 2025c). These sources do not by themselves prove architectural negligence, but they motivate a formal treatment: if the system continuously encounters identity abuse at scale, then the design of identity, reference, and accountability mechanisms is not peripheral but central.

The goal is therefore twofold. First, to formalize the identity and trust problem in a way that makes causal chains explicit and falsifiable. Second, to propose constraint-restoring alternatives that remain compatible with privacy, pseudonymity, and global accessibility, while supplying regulators and researchers with measurable audit targets.

2 Definitions and Framework

This section introduces core terms with explicit definitions and a minimal formal model. Throughout, "Facebook" refers to the Facebook social networking service and its identity, content, and distribution mechanisms, not to the corporate entity as such, except where explicitly stated.

Definition 1 (Entity, Identity, and Representation). *Let E be the set of entities that act on a platform, including individuals, businesses, organizations, and coordinated groups. Let A be the set of platform accounts. Let R be the set of user-facing representational surfaces, including display names, profile images, page titles, and other branding cues. An account $a \in A$ is a platform-side handle for action. A representation $r \in R$ is the interface-level symbol by which users infer referents. An entity $e \in E$ is the real-world or socially meaningful referent of actions.*

Definition 2 (Identity Binding). *An identity binding is a relation $B \subseteq A \times E$ that associates accounts to entities with sufficient stability to support attribution. In a strict regime, B is functional and persistent over time. In a weak regime, B may be non-functional, non-persistent, or partially hidden from users.*

Definition 3 (Namespace Integrity). *A system has namespace integrity at the representational layer if the mapping from representations to entities is injective in the relevant operational domain. Formally, if $I : R \to E$ is the user-inferred mapping induced by platform cues, the system has strong namespace integrity when I is close to injective under ordinary use, meaning collisions are rare, detectable, and strongly penalized.*

Definition 4 (Event-Historical Architecture). *An event-historical architecture is an identity design in which an account's history is represented as an append-only sequence of publicly legible events. Let $H(a)$ denote the event history of account a. An event-historical architecture makes salient events—such as name changes, page category shifts, enforcement actions, and ownership transfers—legible to other users in a structured form that supports auditing and inference.*

Definition 5 (Engagement Extraction). *Let U be the set of users and C the set of content items. A distribution mechanism selects a feed exposure function $D : U \times T \to \mathcal{P}(C)$ over time T. The system performs engagement extraction when D is optimized primarily for measurable interactions (clicks, watch time, comments, shares) rather than for reliability of reference, provenance, or user-aligned intent. This framing aligns with prior analyses of platform incentives under attention-based business models (Wu 2016; Zuboff 2019).*

Definition 6 (Trust Entropy). *Let S denote a measure of uncertainty experienced by users about the referent and reliability of encountered content and accounts. Trust entropy is the growth of S over time induced by ambiguous identity, deceptive representation, and insufficient provenance. The term is used as an analogy grounded in information theory: when disambiguating information is removed or made inaccessible, uncertainty increases and verification costs rise (Shannon 1948).*

3 Assumptions and Scope

The argument relies on explicit assumptions about what a trust-bearing platform must provide.

Assumption 1 (Attribution Requirement). *A trust-bearing social system must support attribution with bounded error under ordinary use. Concretely, users must be able to determine whether an account or page claiming to represent an entity plausibly does so without requiring external investigation in the typical case.*

Assumption 2 (Persistence Requirement). *A trust-bearing social system must supply identity persistence sufficient to impose durable consequences on repeated abuse. If adversaries can cheaply reset identity histories while retaining the ability to reach targets, deterrence collapses.*

Assumption 3 (Collective Defensibility). *A trust-bearing system must permit communities to share defensive information at scale, including block signals, scam signatures, and impersonation markers, subject to abuse-resistant controls. Otherwise, attackers learn globally while defenders learn locally.*

These assumptions are not claims that anonymity or pseudonymity must be eliminated. They are claims that reference and accountability must be preserved within the system's own semantics, much as naming systems in distributed computing require stable namespaces to support resolution and coordination (Mockapetris 1987a; Mockapetris 1987b).

4 Axioms of Trust-Bearing Identity Systems

We now state axioms in a numbered form used throughout the manuscript. They are descriptive constraints: if they are violated, predictable failure modes follow.

Axiom 1 (A1: Identity Persistence). *For each account $a \in A$, the platform must maintain a persistent binding to its event history $H(a)$ such that significant identity transitions are not erasable in the ordinary course of use.*

Axiom 2 (A2: Namespace Uniqueness at the Interface). *The platform must minimize representational collisions in R in the domains where users make high-stakes decisions, and must provide strong disambiguators where collisions occur.*

Axiom 3 (A3: Historical Legibility). *For users to calibrate trust, the platform must expose sufficient structured elements of $H(a)$ to allow attribution and risk inference without requiring external tools.*

Axiom 4 (A4: Constraint Precedence). *Optimization objectives such as engagement, growth, and monetization must be subordinated to identity and provenance constraints. Optimization applied before constraint satisfaction amplifies adversarial strategies.*

Axiom 5 (A5: Entropy Boundedness). *The platform must actively bound trust entropy by detecting and suppressing recurring identity abuse patterns with durable interventions, not solely surface deletion.*

Axiom 6 (A6: Collective Defensibility). *The platform must permit scalable, privacy-respecting sharing of defensive state, so that protection can propagate at least as efficiently as attacks.*

5 Contributions and Roadmap

The remainder of the manuscript proceeds in seven stages. First, it formalizes the interface-level identity mapping $I : R \to E$ and demonstrates why non-injectivity at scale is not merely confusing but structurally exploitable. Second, it models enforcement without durable memory as a near-Markov process, clarifying why repeated scams recur under local deletion regimes. Third, it connects identity ambiguity to information-theoretic uncertainty, introducing quantitative measures and audit metrics grounded in Shannon-style reasoning (Shannon 1948). Fourth, it strengthens the thermodynamic analogy by defining a microstate interpretation for trust and stating where the analogy does and does not apply. Fifth, it develops an economic model in which scamming can become a Nash equilibrium under low-cost reset and engagement-coupled distribution, aligning with classic information-asymmetry results (Akerlof 1970). Sixth, it builds an empirical program: platform transparency metrics, documented case-study archetypes, and comparative analysis across identity regimes. Finally, it provides an implementation roadmap and a policy framework consistent with existing legal structures for transparency and systemic risk management in platform governance (Regulation (EU) 2022/2065; Regulation (EU) 2016/679).

In computer systems terms, the manuscript treats identity as a naming problem and trust as a coordination problem. Naming systems such as DNS exist precisely because global coordination requires stable resolution of symbols to referents, and because failure to resolve names reliably produces systemic vulnerabilities (Mockapetris 1987a; Mockapetris 1987b). Similarly, distributed systems literature emphasizes that adversarial environments require designs resilient to Byzantine behavior, meaning behavior that is arbitrary, strategic, and difficult to distinguish from honest operation (Lamport, Shostak, and Pease 1982; Castro and Liskov 1999). The central claim here is that Facebook-scale identity and distribution should be analyzed under comparable assumptions: the environment is adversarial, the incentives are extractive, and the failure of naming is not a minor UI defect but a governance failure with measurable externalities.

6 Formal Model of Identity Mapping and Non-Injectivity

We now formalize the identity problem more precisely. Let R denote the set of user-visible representations and E the set of entities, as defined earlier. Users implicitly construct an inference function

$$I_u : R \to \mathcal{P}(E),$$

where $\mathcal{P}(E)$ denotes the power set of entities, reflecting uncertainty. In an ideal trust-bearing system, $I_u(r)$ is either a singleton or a small, clearly disambiguated set for representations used in consequential contexts.

On Facebook-scale systems, I_u is systematically non-injective. Distinct entities may share identical or near-identical representations, including names, profile images, page titles, and content style. The platform does not enforce a constraint ensuring that $|I_u(r)| \approx 1$ for representations that are algorithmically amplified or monetized. Instead, collisions are common, persistent, and inexpensive to generate. This property can be demonstrated empirically by sampling pages named after known businesses or public figures and observing the frequency of indistinguishable or minimally distinguishable replicas, a phenomenon documented repeatedly in investigative reporting and internal disclosures (Horwitz and Seetharaman 2021; Reuters 2025a).

Proposition 1. *If the representational mapping I_u is non-injective with low collision penalty, then impersonation-based fraud is a rational strategy for adversarial entities.*

The justification follows directly from incentive structure. When creating a colliding representation incurs low cost and yields access to an audience that cannot reliably disambiguate referents, the expected payoff of impersonation dominates honest signaling. This is a direct extension of adverse selection results: honest entities incur higher signaling costs while deceptive entities exploit ambiguity, driving equilibrium behavior toward imitation (Akerlof 1970).

This property is not merely theoretical. Meta's own transparency reports acknowledge the continual removal of large volumes of fake or impersonating accounts, implying that the cost of entry remains low relative to the cost of detection (Meta Transparency Centre 2025). Persistent high removal volume is therefore evidence not of success but of a stable adversarial equilibrium.

7 Enforcement as a Memoryless Process

We now examine enforcement dynamics. Let S_t denote the global system state at time t, including the set of active accounts and their histories. Let E_t denote the enforcement action applied at time t, typically deletion or disabling of specific accounts. In the absence of durable identity persistence, enforcement can be modeled as a near-Markov process in which the next state depends primarily on the current surface state rather than on accumulated historical information.

Formally, if enforcement does not propagate constraints to future identity instantiations, then

$$P(S_{t+1} \mid S_t, E_t) \approx P(S_{t+1} \mid S_t),$$

because the removal of an account erases information rather than encoding it into future constraints. The system therefore forgets faster than adversaries learn. Attackers, by contrast, retain memory across attempts and refine tactics accordingly, producing a learning asymmetry.

Proposition 2. *In a system where enforcement lacks historical persistence, repeated adversarial behavior converges toward improved effectiveness over time.*

This proposition aligns with empirical observations in adversarial machine learning and sparse representation theory, where attackers exploit brittle classifiers and iterate faster than defenses that lack memory or shared state (Elad 2010). It also aligns with whistleblower accounts indicating that scam networks repeatedly reconstitute after takedowns, often with higher conversion efficiency (Haugen 2021; Reuters 2025b).

8 Information Loss and Trust Entropy

The non-injective identity mapping and memoryless enforcement regime jointly destroy disambiguating information. This can be framed using Shannon entropy. Let X be a random variable representing the true entity behind a representation encountered by a user. Let R be the observed representation. The conditional entropy $H(X \mid R)$ measures user uncertainty about the referent.

In a system with strong namespace integrity and historical legibility, $H(X \mid R)$ is low for consequential interactions. On Facebook-scale systems, representational collisions and hidden histories increase $H(X \mid R)$ systematically. Each removal of historical context or suppression of provenance cues increases uncertainty, forcing users to rely on heuristics such as engagement metrics or superficial cues.

Proposition 3. *Architectures that suppress or erase identity history increase $H(X \mid R)$ monotonically over time, absent compensating disambiguation mechanisms.*

This increase in conditional entropy corresponds to what we previously termed trust entropy. While the thermodynamic analogy has limits, the information-theoretic grounding is precise: destroying information increases uncertainty and raises the cost of verification (Shannon 1948). When verification costs exceed expected benefit for ordinary users, rational disengagement or heuristic trust becomes the norm.

9 Relation to Naming Systems and Distributed Computing

The identity failures observed here mirror well-studied problems in computer systems. Naming systems such as the Domain Name System exist precisely to enforce global uniqueness and resolvability of identifiers under adversarial conditions (Mockapetris 1987a; Mockapetris 1987b). Public-key infrastructures similarly bind identities to cryptographic material to prevent impersonation at scale. These systems accept friction and overhead because the alternative is systemic ambiguity.

Distributed systems research further demonstrates that in adversarial environments, assumptions of honesty are insufficient. Byzantine fault tolerance models assume arbitrary, deceptive behavior and design consensus mechanisms accordingly (Lamport, Shostak, and Pease 1982; Castro and Liskov 1999). Facebook-scale identity systems, by contrast, are designed as if adversarial behavior were peripheral rather than central, despite overwhelming evidence to the contrary.

This contrast is instructive. In safety-critical domains, identity and history are not optimized away for convenience. They are preserved precisely because coordination and trust depend on them. The argument advanced here is that social platforms of comparable scale and impact should be evaluated under similar design expectations.

10 Falsifiable Predictions

The formal framework yields testable predictions. If identity persistence and namespace integrity are strengthened in a controlled environment, then scam recurrence rates should decrease superlinearly relative to enforcement effort. If event-historical cues are made legible to users, then reliance on engagement metrics as trust proxies should decline, measurable through changes in click-through and reporting patterns. Conversely, further suppression of historical cues without constraint restoration should correlate with rising scam conversion efficiency even if takedown volume increases.

These predictions can be evaluated using platform-provided data, independent audits, or regulatory-mandated disclosures. Their falsifiability distinguishes the argument from purely normative critiques and grounds it in measurable system behavior.

11 Transition to Economic and Thermodynamic Analysis

Having established the formal identity and information-theoretic structure, the next section develops the thermodynamic analogy with greater precision and integrates it with an explicit economic model. The goal is not metaphorical flourish but analytical leverage: to show how trust degradation behaves like an externalized cost and why, under current incentive structures, it is rational for platforms to underinvest in constraint restoration even when harms are obvious.

12 Thermodynamic Framing of Trust Degradation

We now deepen the thermodynamic analogy introduced earlier and clarify both its utility and its limits. The analogy is not intended to imply that social systems obey physical laws, but rather that certain information-theoretic regularities governing uncertainty, dissipation, and equilibrium apply when large populations interact under constrained attention and imperfect information.

Let Ω denote the space of microstates corresponding to possible configurations of identity, representation, and historical information available to users at a given time. Each microstate encodes which disambiguating cues are visible, which histories are suppressed, and which representations collide. A macrostate corresponds to an aggregate level of user uncertainty about identity and provenance. Trust entropy, denoted S_T, is defined as a monotonic function of the average conditional entropy $H(X \mid R)$ across typical interactions, where X is the true entity and R the observed representation.

In architectures with strong namespace integrity and historical legibility, microstates that preserve disambiguating information dominate. In architectures that suppress history and permit representation collisions, the system explores a larger portion of Ω associated with ambiguity. As a result, S_T increases. This increase is not random. It is driven by repeated design decisions that favor short-term engagement over information preservation.

Remark 1. *The thermodynamic analogy breaks down if interpreted literally. There is no conserved energy analogue, and human agents adapt strategically. The analogy is useful only insofar as it captures the directional tendency of uncertainty to increase when disambiguating information is removed faster than it is replenished.*

What makes the analogy powerful is that it highlights irreversibility. Once trust entropy rises past certain thresholds, restoring low-entropy states becomes costly. Users disengage, rely on heuristics, or abandon the platform for high-stakes interactions. This resembles critical transitions in complex systems, where gradual parameter changes produce abrupt qualitative shifts (Scheffer et al. 2009). In this framing, scam prevalence and metric laundering are not merely bad outcomes; they are symptoms of a system approaching or surpassing a trust phase transition.

13 Economic Model of Scam Equilibria

The thermodynamic perspective can be coupled with a simple economic model to explain why scamming becomes stable under current architectures. Consider a repeated game with three classes of actors: the platform P, honest users H, and adversarial scammers A. The platform chooses design parameters governing identity persistence and distribution optimization. Honest users choose whether to engage, disengage, or rely on heuristics. Adversaries choose between honest signaling and impersonation-based deception.

If identity persistence is weak and representational collisions are cheap, the payoff to impersonation for A exceeds that of honest signaling, given comparable reach and lower cost. For H, the cost of verification rises as trust entropy increases, making heuristic trust or disengagement rational. For P, engagement-optimized distribution yields revenue proportional to interaction volume, while the cost of scam mitigation is internalized only partially, as many harms are borne by users and society at large.

Proposition 4. *Under weak identity persistence and engagement-coupled distribution, impersonation-based scamming constitutes a Nash equilibrium.*

The justification follows from standard game-theoretic reasoning. Given the platform's incentive to maximize engagement and the absence of durable penalties for identity abuse, no single actor benefits from unilateral deviation. Honest users cannot easily enforce constraints, scammers profit from impersonation, and the platform continues to extract value from interaction volume. This mirrors classic externality problems in markets with information asymmetry, where private incentives diverge from social welfare (Akerlof 1970).

Quantifying the externality is challenging but conceptually straightforward. The social cost includes lost consumer surplus from fraud, increased verification costs, erosion of institutional trust, and diversion of attention from productive activities. These costs are not fully reflected in platform decision-making, leading to underinvestment in constraint-restoring infrastructure. The result is an equilibrium that is privately stable but socially inefficient.

14 Empirical Signals and Quantitative Trends

While comprehensive data access is limited, several quantitative signals support the structural analysis. Meta's transparency reporting consistently acknowledges the removal of large volumes of fake accounts each quarter, often numbering in the billions annually across its family of products (Meta Transparency Centre 2025). High removal volume combined with persistent prevalence estimates implies rapid regeneration, consistent with low-cost identity reset.

Investigative reporting has further documented scam advertising networks that repeatedly reappear after enforcement, sometimes within days, often using near-identical creative assets and page structures (Reuters 2025a; Reuters 2025b). Internal documents cited in whistleblower testimony suggest that the company has long been aware of the revenue generated by such activity and of trade-offs between enforcement aggressiveness and business metrics (Haugen 2021).

Comparative signals also matter. Platforms with stronger identity persistence or professional context constraints, such as LinkedIn or GitHub, report lower prevalence of impersonation in high-stakes interactions, despite operating at substantial scale. This does not imply that such platforms are immune to abuse, but it suggests that namespace integrity and event history materially affect equilibrium behavior. Conversely, platforms emphasizing frictionless account creation and algorithmic amplification exhibit similar scam patterns, indicating that the observed dynamics are not idiosyncratic to a single company.

15 Case Archetypes of Scam Evolution

Across platforms and reporting sources, several recurring scam archetypes illustrate adaptive behavior under weak identity constraints. Impersonation of known brands or public figures exploits representational collisions. Account takeover or lookalike pages leverage inherited trust. Monetization scams exploit aspirational labor incentives. Each archetype evolves in response to enforcement, modifying surface features while preserving underlying strategy.

What unifies these cases is not their specific content but their reliance on low-cost identity reset and user-facing ambiguity. Enforcement actions that target individual accounts remove instances without altering the generative process. From a systems perspective, this is analogous to treating symptoms while leaving the causal mechanism intact.

16 Transition to Measurement and Policy

The preceding analysis motivates a shift from diagnosis to measurement and intervention. If trust degradation is structural, then effective governance requires metrics that capture identity integrity and scam amplification directly, rather than relying solely on takedown counts or engagement statistics. The next sections therefore propose concrete measurement frameworks, audit mechanisms, and an implementation roadmap for constraint-restoring architectures, followed by a legal and policy analysis grounded in existing regulatory regimes.

17 Measurement and Detection of Identity and Trust Failure

If trust degradation is a structural property rather than an incidental outcome, then it must be measurable in ways that go beyond anecdotal reporting or raw enforcement counts. The dominant metrics currently emphasized by platforms, such as the number of accounts removed or pieces of content taken down, are insufficient because they do not distinguish between suppression of symptoms and correction of causes. A system that removes large numbers of fraudulent accounts while permitting rapid regeneration may appear active while remaining ineffective.

A more appropriate measurement framework begins with identity integrity. One candidate metric is the representational collision rate, defined as the proportion of user-visible representations that plausibly map to more than one entity in consequential contexts. This can be operationalized by sampling pages or accounts claiming affiliation with known organizations or public figures and measuring the frequency with which indistinguishable or weakly distinguishable replicas appear. A second metric is identity half-life, defined as the expected time for an adversarial entity to regain comparable reach after enforcement. Short half-lives indicate weak persistence constraints and high adversarial adaptability.

Trust entropy can be approximated empirically by measuring user uncertainty proxies. These include reliance on engagement metrics in decision-making, increased reporting rates without corresponding decreases in scam prevalence, and declining interaction with organic content in favor of paid or platform-verified sources. Longitudinal changes in these proxies can be correlated with design interventions to test causal hypotheses. For example, if adding event-historical cues reduces reliance on likes or follower counts as credibility signals, this would support the claim that historical legibility lowers conditional entropy.

Detection experiments should therefore focus on structural interventions rather than classifier tuning. Controlled rollouts that increase identity persistence for a subset of accounts, or that expose limited historical context to users, can be evaluated for effects on scam recurrence and user behavior. Crucially, such experiments must be designed to measure not only immediate outcomes but also adversarial adaptation over time. Short-term reductions that vanish after attackers adjust do not constitute success.

18 Implementation Roadmap for Constraint-Restoring Architectures

Restoring identity constraints at Facebook scale is not a single engineering task but a phased institutional transformation. Some changes are immediately feasible, while others require careful migration to avoid collateral harm.

The most urgent interventions are those that reduce adversarial learning asymmetry. Shared defensive mechanisms, such as opt-in block list sharing or scam signature propagation, can be introduced incrementally without altering account semantics. These measures primarily benefit users and moderators while imposing minimal friction on legitimate participation.

A second phase involves event-historical augmentation. Rather than imposing rigid real-name policies, platforms can attach structured, append-only metadata to accounts and pages. Examples include visible indicators of recent name changes, page category transitions, ownership transfers, and prior enforcement actions, presented in a manner that preserves privacy while restoring context. Backward compatibility can be managed by introducing history accumulation prospectively, while gradually integrating legacy data where reliable records exist.

The most challenging phase concerns namespace integrity. Enforcing stronger disambiguation for high-risk representational domains, such as business pages, political entities, or monetized accounts, requires accepting some friction. This may include verification steps, delayed amplification for newly created representations, or mandatory disambiguators when collisions occur. The cost of such measures is reduced accessibility and slower onboarding for some users. The benefit is a reduction in systemic impersonation risk. This trade-off must be evaluated explicitly rather than obscured by engagement metrics.

Implementation costs are nontrivial but bounded. Large platforms already operate extensive trust and safety infrastructure and maintain detailed internal logs. The primary expense lies not in computation but in organizational willingness to prioritize constraint enforcement over short-term growth. The benefits, while harder to quantify, include reduced fraud losses, improved user retention in high-stakes domains, and lower regulatory risk.

19 Comparative Platform Analysis

Examining alternative identity regimes clarifies which design choices matter. Platforms such as LinkedIn operate under a professional identity norm that discourages frequent identity resets and provides contextual signals tied to employment history. While abuse persists, impersonation in consequential professional interactions is comparatively rarer. GitHub enforces persistent usernames linked to public contribution histories, creating strong reputational inertia even under pseudonymity. Identity is not necessarily real-name, but it is durable and historically legible.

Decentralized platforms such as Mastodon and Bluesky illustrate different trade-offs. Mastodon's federated architecture allows communities to enforce local identity norms and share block information across instances, enhancing collective defense. However, federation introduces coordination challenges and uneven enforcement. Bluesky's decentralized identifier system separates identity from hosting, potentially enabling stronger persistence, but practical outcomes remain under evaluation.

Historical precedents also matter. Early forums and bulletin board systems often enforced persistent handles and visible moderation logs, creating reputational continuity despite anonymity. While these systems lacked scale, they demonstrate that persistent pseudonymity can support trust without requiring real-world identity disclosure.

The comparative evidence suggests that identity persistence and historical legibility, rather than real-name enforcement, are the critical variables. Systems that preserve these properties exhibit lower impersonation risk and slower adversarial adaptation, even when operating under diverse governance models.

20 Legal and Policy Framework

Existing legal frameworks already recognize the systemic risks posed by opaque, large-scale intermediaries. The European Union's Digital Services Act establishes obligations for very large online platforms to assess and mitigate systemic risks, including the dissemination of illegal content and negative effects on fundamental rights. While the Act does not mandate specific identity architectures, its emphasis on risk assessment, transparency, and auditability aligns with the constraint-restoring approach advocated here.

Data protection regimes such as the General Data Protection Regulation impose requirements for purpose limitation, data minimization, and user rights. Event-historical architectures must therefore be designed to balance accountability with privacy. This tension is not unique to social platforms; financial and telecommunications systems have long reconciled persistent identity with regulated access and oversight.

In jurisdictions governed by intermediary liability protections, such as those historically associated with Section 230 in the United States, the argument advanced here reframes responsibility. The claim is not that platforms are liable for all user behavior, but that when architectural choices foreseeably amplify fraud, regulatory scrutiny of those choices is appropriate. This parallels regulatory approaches in other industries where design decisions that externalize risk trigger obligations to mitigate.

Policy proposals consistent with existing frameworks include mandatory disclosure of identity reset rates, standardized reporting of scam recurrence, and independent audits of representational collision rates. These measures do not prescribe implementation details but create accountability incentives aligned with public interest.

21 Addressing Counterarguments and Trade-Offs

Several counterarguments merit serious engagement. One concerns scale. Maintaining unique namespaces and persistent histories across billions of users is technically challenging. This is true. However, other global systems, including DNS and payment networks, operate under similarly adversarial conditions by accepting overhead and governance complexity as necessary costs. Scale does not eliminate the need for constraints; it magnifies it.

Another counterargument concerns accessibility and inclusion. Stronger identity constraints may exclude users who lack documentation or who rely on anonymity for safety. This risk is real and must be addressed through design choices that separate persistence from real-world identification. Pseudonymous but persistent identities, graduated trust levels, and context-specific constraints offer ways to balance inclusion with accountability.

A further explanation is organizational inertia or coordination failure rather than intentional design. Large organizations may struggle to align incentives across teams, leading to suboptimal outcomes without malicious intent. This possibility does not negate the structural analysis. Whether harms arise from intent or inertia, the causal mechanisms remain and require correction. Regulatory and governance interventions are designed precisely for cases where private incentives fail to align with social welfare.

22 Conclusion

The central claim of this manuscript is that trust degradation and scam amplification on Facebook-scale platforms are structurally produced rather than incidentally hosted. By relaxing identity persistence, permitting representational collisions, and optimizing distribution for engagement, platforms create an environment in which impersonation-based fraud is a rational and stable strategy. Formal modeling, information-theoretic analysis, economic reasoning, and empirical signals all converge on this conclusion.

These failures differ from historical infrastructure disasters whose harms were uncertain at deployment. The mechanisms at issue here are conceptually elementary and have been visible for years. That they persist reflects prioritization choices rather than epistemic limits. At the same time, the existence of harm does not imply the absence of solutions. Constraint-restoring architectures are feasible, measurable, and compatible with privacy when designed carefully.

At planetary scale, digital identity infrastructure shapes economic behavior, civic trust, and the allocation of human attention. When that infrastructure predictably amplifies deception, responsibility cannot be displaced onto users or attackers alone. The appropriate response is not resignation but redesign, guided by explicit constraints, empirical audit, and governance commensurate with impact.

A Addendum: Institutional Rebranding and Legitimacy Transfer

The identity dynamics described throughout this work are not confined to user behavior. They are practiced at the institutional level by platform operators themselves. The corporate rebranding of Facebook to Meta illustrates the same non-injective identity mapping observed in scam behavior, whereby a persistent underlying entity adopts a new public identifier to shed accumulated reputational debt while preserving operational continuity.

This strategy is reinforced through association with high-prestige scientific and humanitarian initiatives, such as the Chan Zuckerberg Biohub, a nonprofit research organization founded by Mark Zuckerberg and Priscilla Chan to pursue ambitious goals in AI-powered biology. The value of such work is not in question. What is structurally relevant is the transfer of legitimacy across domains. Positive identity in one arena is allowed to offset negative identity in another despite the absence of causal separation. This is identity remix at institutional scale.

That these techniques are employed openly underscores the foreseeability of the harms discussed in this paper. The organization demonstrates a sophisticated understanding of how naming, affiliation, and legitimacy function as transferable assets. The failure to enforce analogous constraints within its own platforms therefore cannot plausibly be attributed to ignorance. It reflects selective application of identity integrity, enforced where advantageous and relaxed where profitable.

References

[1] G. A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

[2] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[3] A.-L. Barabási. *Network Science.* Cambridge University Press, 2016.

[4] J. Bridle. *New Dark Age: Technology and the End of the Future.* Verso, 2018.

[5] D. T. Campbell. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90, 1979.

[6] C. Doctorow. *The Internet Con: How to Seize the Means of Computation.* Verso, 2023.

[7] M. Elad. *Sparse and Redundant Representations.* Springer, 2010.

[8] L. Floridi. *The Fourth Revolution.* Oxford University Press, 2014.

[9] C. A. E. Goodhart. Problems of monetary management. Reserve Bank of Australia, 1975.

[10] F. Haugen. Testimony before the U.S. Senate Subcommittee on Consumer Protection. 2021.

[11] J. Horwitz and D. Seetharaman. The Facebook files. *The Wall Street Journal*, 2021.

[12] J. Lanier. *Ten Arguments for Deleting Your Social Media Accounts Right Now.* Henry Holt, 2018.

[13] L. Lessig. *Code and Other Laws of Cyberspace.* Basic Books, 1999.

[14] M. Scheffer et al. Early-warning signals for critical transitions. *Nature*, 461:53–59, 2009.

[15] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[16] H. A. Simon. Designing organizations for an information-rich world. Johns Hopkins University Press, 1971.

[17] Z. Tufekci. Algorithmic harms beyond Facebook and Google. *Colorado Technology Law Journal*, 2015.

[18] Y. Varoufakis. *Technofeudalism.* Melville House, 2023.

[19] T. Wu. *The Attention Merchants.* Knopf, 2016.

[20] S. Zuboff. *The Age of Surveillance Capitalism.* PublicAffairs, 2019.