Journal of Institutional Economics (2026), **22**, e32, 1–23
doi:10.1017/S1744137426100587

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Verified pseudonymity as governance technology: decentralized identity, soulbound tokens, and online discourse

Christos A. Makridis[1,2,3]

[1]Stanford University, USA, [2]Arizona State University, W. P. Carey School of Business, Business Administration, USA and [3]University of Nicosia, Institute for the Future, AGC Towers, 28th October 24, Nicosia, Cyprus
Email: cmakridi@asu.edu

(Received 10 March 2026; revised 25 April 2026; accepted 25 April 2026)

Abstract

Cheap, disposable online identities make abuse easier to externalize. Users can harass, evade bans, amplify content through fake accounts, or abandon a damaged reputation at low cost, while other users, moderators, and platforms bear the consequences. This paper examines verified pseudonymity as an institutional response to that problem. First, I model online communities as club-governed informational commons in which incivility degrades the shared environment and raises enforcement costs. The model shows that conduct can improve when sanctions attach to a persistent pseudonymous identity and when users have future access, reputation, or governance rights at stake. Second, I compare verified pseudonymity with open pseudonymity, real-name mandates, centralized know-your-customer verification, algorithmic moderation, and no intervention. Decentralized identifiers, verifiable credentials, proof of personhood, and non-transferable standing credentials matter because they can separate authentication from public identification. Third, I add a community-currency layer that separates access to scarce attention from governance rights. The result is a governance framework in which accountability depends less on public naming than on durable standing, credible sanctions, and reusable privacy-preserving credentials.

Keywords: anonymity; decentralized identity; online incivility; polarization; soulbound tokens

JEL classifications: D83; D85; L86; O33; K24

Introduction

The rise of online polarization and toxicity is a pressing concern for democracies and platform governance. Extreme incivility has proliferated on social media, often alongside growing ideological echo chambers, with consequences for mental health, civic trust, and democratic participation. A large body of research has pointed to user anonymity and fake accounts as enabling factors.[1] When people hide behind pseudonyms or multiple unlinked accounts, they may feel emboldened to engage in antisocial behaviour without accountability – a phenomenon known as the online disinhibition effect (Suler, 2004). In addition, malicious actors can exploit the openness of online networks by creating numerous fake personas to spread propaganda and inflame divisions. This has led to the intuitive claim that tying online profiles to *real-world verified identities* could improve online discourse by injecting

[1]See, for instance, Suler (2004); Santana (2014); Rösner and Krämer (2016); and Shao *et al.* (2018).

© The Author(s), 2026. Published by Cambridge University Press on behalf of Millennium Economics Ltd. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

accountability and deterring bad actors. 'Skin in the game' – i.e., reputation or identity at stake when posting – could create stronger incentives to dampen otherwise harmful allegations.

Does identity verification offer a step towards solving online toxicity and polarization? On one hand, decades of research in computer-mediated communication indicate that anonymity can decrease social inhibitions and empathy, leading to nastier interactions (Suler, 2004). Studies of news comment sections found that users posting under real names or verified identities were significantly more civil on average than those posting anonymously (Coe *et al.*, 2014; Santana, 2014). On the other hand, real-name policies have drawbacks: they may discourage participation if users fear privacy risks or retribution for their views and they do not necessarily ameliorate deeper issues of ideological polarization or misinformation. For instance, Facebook has long had a real-name rule and yet polarized content remains on it (Bakshy *et al.*, 2015). Identity confirmation might tone down the incivility of discourse without changing the division in viewpoints.

This paper reframes the problem in institutional-economic terms and develops a governance architecture designed to raise the cost of antisocial conduct without requiring universal public identification. In short, this paper suggests that online incivility, harassment, ban evasion, and inauthentic participation are best understood as externality problems rooted in cheap, disposable identities – not as inevitable features of pseudonymous speech. When an identity can be abandoned at negligible cost, the intertemporal discipline that ordinarily restrains opportunism collapses: users capture the private gains from abuse while externalizing the resulting deterioration of discourse quality, moderator capacity, and participation incentives (Celuch *et al.*, 2024; Coe *et al.*, 2014; Friedman and Resnick, 2001). The question is how to restore that discipline without importing the well-documented costs of real-name mandates and centralized verification.

I begin by developing the economics of incivility through a stylized model. Online communities are modelled as clubs governing access to an informational commons (Buchanan, 1965; Ostrom, 1990): excludable at the boundary through admission and credential rules, but congestible and degradable from within as incivility rises. Building on Friedman and Resnick (2001), when identity-reset costs are low, communities face a distorted equilibrium: they must either tolerate more abuse or impose diffuse suspicion on newcomers by withholding trust and standing until dues have been paid. Institutions that raise the cost of identity reset and increase the continuation value of good standing relax that tradeoff. Identity governance affects conduct – harassment, manipulation, Sybil attacks – by creating a 'durable identity' because it alters the expected penalty for misbehaviour. It should not be expected to dissolve ideological sorting or overcome the informational cascades, homophily, and selective exposure that independently drive polarization (Bail *et al.*, 2018; Bakshy *et al.*, 2015; Sunstein and Hastie, 2015).

Next, I evaluate the relevant governance alternatives as comparative-institutional bundles, varying across admission costs, privacy costs, monitoring costs, sanction credibility, portability of standing, and concentration of control. Doing nothing preserves openness but absorbs the full cost of spam and Sybil attacks. Real-name rules raise the cost of some misconduct but create chilling effects and centralized databases of personal information. Using an illustrative example from South Korea, I explain how a mandatory real-name verification regime – struck down as unconstitutional in 2012 – reduced participation more clearly than it improved discourse quality (Cho and Kwon, 2015; Leitner, 2009). Centralized platform KYC is more flexible but fragments identity across firms and concentrates gatekeeping power. Algorithmic moderation scales but cannot eliminate the need for human judgment, appeals, and institutional safeguards (Davidson, 2025; Howell and Potgieter, 2021). This comparative analysis motivates the architecture I propose.

Next, I develop a mechanism design for verified pseudonymity as a hybrid governance architecture. Four layers – proof-of-personhood, decentralized identifiers, verifiable credentials (including soul-bound-token-style non-transferable standing), and platform integration with graduated sanctioning and appeals – are designed to raise the cost of identity reset and increase the continuation value of good standing while preserving pseudonymous public participation. The key is the non-transferable continuation value it attaches to compliant behaviour: access to verified channels, governance eligibility, and trust signals that cannot be sold or transferred to a fresh account (Ohlhaver *et al.*, 2022). Privacy is

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

maintained through selective disclosure: participants prove uniqueness or credential-holding without revealing legal names, using decentralized identifiers and the verifiable-credentials model to separate verification from publicity.

The architecture is then extended with a fifth layer that addresses a limitation identity governance alone cannot solve: the allocation of scarce community resources. Building on Ohlhaver's (2025) Plural Community Asset Resource Exchange (PCARE) model, I introduce a community-currency system that prices attention separately from influence. A transferable attention token pays for bandwidth – posting in high-reach channels, initiating appeals, requesting human review – while a non-transferable influence stake governs policy rights with diminishing returns through square-root weighting. This separation prevents liquid wealth from converting directly into governance power and creates context-specific pricing across community boundaries that makes spam, coordinated manipulation, and resource exhaustion more expensive. The value becomes especially clear in the emerging agentic AI environment. As software agents become cheap to deploy, the governance problem extends beyond one human using many accounts to one principal directing many agents. Each request in the proposed system carries three signals – an identity credential that authorizes it, a token spend that pays for attention, and staked influence that can be lost if rules are broken – ensuring that delegation does not circumvent accountability.

This paper contributes to a large literature in institutional economics on clubs, commons, reputation, and polycentric governance in digital environments by treating online discourse as a club-governed informational commons whose quality depends on boundary rules, information rules, and sanction rules (Buchanan, 1965; Harris, 2018; Ostrom, 1990; Safner, 2016). My stylized model shows that incivility, harassment, ban evasion, and inauthentic participation create externalities that become more severe when identities are cheap to create, abandon, and replicate. Recasting the problem in those terms shifts the analysis towards a more precise institutional question: how to make exclusion, durable standing, and repeated interaction credible in a privacy-preserving and decentralized fashion – that is, without requiring universal public identification.

The main comparative claim in this paper is that verified pseudonymity can raise sanction credibility and reduce identity-reset arbitrage without imposing the full privacy and participation costs of real-name rules or centralized know-your-customer (KYC) laws. Rather than offering a general defence of identity verification, it advances a narrower claim. Privacy-preserving verified pseudonymity is more likely to reduce incivility, harassment, Sybil attacks, ban evasion, and some forms of inauthentic participation than to reduce ideological polarization, which is also shaped by homophily, selective exposure, informational cascades, and conflict-sustaining institutions (Bakshy *et al.*, 2015; Bail *et al.*, 2018; Mildenberger, 2018; Mildenberger and Tingley, 2019; Sunstein and Hastie, 2015). The model shows that conduct improves when users have more to lose from sanction and less reason to exit, which is why verified pseudonymity can dominate real-name or centralized verification regimes that raise deterrence by also raising participation costs. Blunt verification regimes can reduce participation without producing commensurate gains in discourse quality, as South Korean's real-name mandate suggests (Leitner, 2009; Constitutional Court of Korea, 2012; Cho and Kwon, 2015), and why purely algorithmic moderation remains constrained by incomplete contracting, context, and the continuing need for review and appeals (Davidson, 2025; Howell and Potgieter, 2021).

Finally, this paper contributes to the institutional-economics literature on blockchain and digital organization by interpreting decentralized identifiers, proof-of-personhood, verifiable credentials, and soulbound-token-style attestations as governance technologies rather than as technical novelties (Alston *et al.*, 2022; Davidson *et al.*, 2018; Frolov, 2021; Ohlhaver *et al.*, 2022; W3C, 2022, 2025). These tools split authority across issuers, users, platforms, and appeals bodies, which can improve privacy and sanction credibility but also creates coordination and accreditation problems. The paper also situates that design within the legal and organizational limits (Devereaux and Peng, 2020; Finck, 2019; Giraudo, 2022; Howell and Potgieter, 2021): decentralization does not eliminate governance, data-protection, or due-process problems; it changes how ecosystem actors engage with them through incentives. On that basis, the paper extends the argument with a fifth layer from Ohlhaver's (2025) PCARE framework,

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

showing how verified communities can separate the price of attention from the cost of influence in ways that bear directly on spam, governance capture, and the emerging problem of agentic AI.

The economics of incivility and accountability

Conceptual underpinnings

Online incivility is not only a moral or psychological concern but it is also an institutional and economic problem because it changes the cost of transacting, speaking, screening, and governing within digital communities (Cho and Kwon, 2015; Coe *et al.*, 2014). Online forums, comment threads, creator platforms, and social networks are not merely venues for expression. They are production environments for information, attention, trust, and cooperation. When harassment, trolling, doxxing threats, brigading, ban evasion, and bot-assisted abuse become common, users devote more resources to defensive behaviour: muting, blocking, screening messages, withholding participation, or exiting altogether. Moderators and platform owners likewise bear higher enforcement and adjudication costs. The result is not only disutility to direct targets, but a thinner and less reliable market for discourse. Valuable participants contribute less, prospective participants face worse expectations about treatment and discussion quality, and the average quality of interaction declines (Celuch *et al.*, 2024; Graf *et al.*, 2017). In advertiser-supported environments, these effects create brand-safety risks and lower monetization (Johnson *et al.*, 2023); in civic, professional, or knowledge communities, they degrade the shared informational environment on which discovery, persuasion, and collaboration depend (Celuch *et al.*, 2024).

In this paper, I refer to reputation to reflect the expected value of 'durable standing': it exists when present conduct can be linked to future opportunities for participation, exchange, and influence.[2] Civility refers to rule-conforming disagreement. Incivility refers to conduct that imposes governance costs on others and degrades the forum, including insults, threats, harassment, doxxing, brigading, ban evasion, and other forms of manipulative or abusive disruption. Verified pseudonymity denotes an arrangement in which a participant is privately authenticated as a unique person or credential-holder while appearing publicly under a stable pseudonym rather than a legal name. The central distinction, then, is not simply anonymity versus publicity. It is between conduct attached to a durable, attributable identity and conduct undertaken through an identity that can be abandoned at low cost. Where identity is disposable, the intertemporal discipline on conduct weakens; where identity is durable, the shadow of the future becomes stronger (Friedman and Resnick, 2001; Suler, 2004). Some users may derive private utility from inflammatory or abusive conduct because it generates attention, amusement, in-group status, or short-run partisan advantage. But those private gains need not exceed the broader costs imposed on other users, moderators, platforms, advertisers, and absent participants who choose not to enter. Incivility is therefore best understood as an externality problem rather than as mere expressive excess.

The governance alternatives are not binary because online discourse spaces are neither pure markets nor pure hierarchies. They are better understood as club-governed informational commons: entry can be restricted, but the shared resource being governed – attention, trust, moderator capacity, and informational quality – remains congestible and vulnerable to degradation from within. This is why the institutional challenge is not only boundary control per se, but also the management of insider conduct. Harris (2018) shows that online communities can sustain cooperation through boundary, position, information, and payoff rules, while Safner (2016) shows that digital commons such as Wikipedia

[2]Reputation, as used here, is a community-specific expectation about future reliability and rule observance built from attributed past behaviour, third-party attestations, and sanction history (Hardy and Norgaard, 2015). It is not the same as popularity, virality, or raw engagement. Civility refers to the manner of disagreement: the extent to which users contest claims without resorting to direct insult, dehumanization, threats, doxxing, harassment, coordinated intimidation, or repeated ban evasion. Incivility describes low-value conduct that imposes conversational and enforcement costs on others even when it yields expressive or strategic utility to the speaker.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

endure through institutional entrepreneurship and governance evolution rather than through openness or protocol design alone. Discourse quality is therefore not an accidental by-product of technology, but rather an outcome of institutional design. Doing nothing preserves maximal openness and avoids direct identity burdens, but it also leaves communities to absorb the full costs of spam, harassment, and Sybil attacks. Blanket real-name rules raise the cost of some misconduct, yet they do so by exposing users to chilling effects, retaliation, and large centralized databases of personal information. Centralized platform KYC is more flexible than public real-name display, but it creates data honeypots, expands platform power over admission and exclusion, and fragments identity across firms. AI or algorithmic moderation scales more readily than manual review, but it faces familiar problems of context, incomplete contracting, and error: automated systems can classify, rank, or suppress content, yet they do not eliminate the need for human judgment, appeals, or institutional safeguards (Davidson, 2025; Howell and Potgieter, 2021). The choice is not between governance and no governance, but among arrangements that distribute privacy burdens, screening costs, and sanctioning power differently.

From an institutional-economics perspective, identity matters because repeated interaction under attributable identity makes exclusion, warning, and reward more credible. In anonymous or weakly authenticated environments, sanctions are easily evaded because harmful behaviour can be shifted to throwaway or duplicate accounts. That weakens ordinary reputational discipline and lowers the expected cost of opportunism. By contrast, durable pseudonymous identity can increase continuity of standing without requiring universal public disclosure of legal names. The logic is familiar from other anonymous or pseudonymous settings: durable reputational records can support self-enforcement even where counterparties are not publicly named (Hardy and Norgaard, 2015), while changes in digital network architecture alter hierarchy and transaction costs more generally (Norgaard *et al.*, 2018). The more plausible margin of improvement is hybrid governance that combines behavioural moderation with privacy-preserving identity and credentialing. Verified pseudonymity is attractive because it seeks to raise the expected cost of misconduct while preserving the participation benefits of pseudonymous public speech.

This framework also clarifies why identity should be expected to affect incivility more directly than polarization. A community can sustain deep ideological disagreement and still function well if disagreement is governed by enforceable norms of conduct. The inefficiency arises not from viewpoint diversity as such, but from low-value behaviour that crowds out exchange, overwhelms moderation capacity, or distorts the apparent composition of the community through fake or disposable accounts. Polarization, by contrast, is also driven by informational cascades, group polarization, selective exposure, and the systematic overweighting of shared over unshared information in collective judgment. Sunstein and Hastie (2015) emphasize that such group-error mechanisms can persist even in deliberative institutions, while Mildenberger (2018) shows that conflictual equilibria can be sustained by conflict-kindling institutions even when participants are not inherently more conflict-prone. Identity-based governance may therefore reduce abuse, manipulation, and certain forms of inauthentic participation by raising the cost of misconduct, but it should not be expected, by itself, to dissolve ideological sorting or reconcile substantive disagreement. Its comparative advantage is governance of conduct, not convergence of beliefs.

Stylized theoretical model

The preceding discussion isolates a narrow mechanism. When identities are cheap to reset, users can retain the private gains from uncivil conduct while externalizing part of the resulting deterioration in discourse quality and governance capacity. The model below formalizes that mechanism. It is intentionally stylized and is designed to explain how verified pseudonymity can reduce incivility, harassment, ban evasion, and related forms of inauthentic participation; it does not attempt to model ideological belief formation or partisan sorting. I build upon Friedman and Resnick's (2001) analysis of 'cheap pseudonyms': when users can costlessly abandon an identity and return under a new one, reputational discipline weakens and communities can be pushed into a distorted equilibrium in which

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

newcomers must 'pay dues' before they are trusted.[3] While the model is admittedly a distant rendering of reality, particularly one as complicated as incivility on online platforms, it nonetheless gives a stylized characterization of the phenomena at play, thereby laying a foundation for future work that may seek to model these dynamics quantitatively.

To place that mechanism in an institutional-economics frame, consider an online community as a club that governs access to an informational commons. The club is excludable at the boundary through admission, credential, and sanction rules, but the resource it protects – trust, attention, moderator capacity, and discourse quality – is congestible and degradable from within. In Buchanan's (1965) sense, the environment is club-like because exclusion is feasible; in Ostrom's (1990) sense, it retains common-pool features because misuse by insiders degrades the environment for other insiders. Let there be a continuum of potential participants of unit mass. Each active participant produces one normalized unit of content. Let $n \in [0,1]$ denote the mass of active participants and let $h \in [0,1]$ denote the share of active contributions that are uncivil, abusive, or otherwise governance-costly. The per-user value of participating in the community is:

$$Q(n,h) = b_0 + \mu n - \lambda h$$

where b_0 is the baseline value of participation, $\mu > 0$ captures the gain from thicker participation and more exchange, and $\lambda > 0$ captures the degradation of the shared discourse environment as incivility rises. The term μn is the club-good side of the model: more participation can raise value by thickening discussion, increasing matching possibilities, and making the community more informative. The term λh is the commons side: as incivility rises, the shared environment deteriorates for everyone using it.

Now consider entry. A potential entrant i has idiosyncratic participation cost c_i. Entering also requires bearing verification or onboarding friction $\tau \geq 0$, which can be interpreted broadly to include time, compliance burden, privacy concern, or wallet-management cost. Finally, entering gives the user an expected continuation value R_0, which is the option value of joining the club as a newcomer with limited standing. User i enters if:

$$Q(n,h) - \tau + R_0 - c_i \geq 0$$

Assume $c_i \sim U[0,1]$. Then, for an interior solution, the mass of entrants is simply the probability that the participation cost lies below the entry threshold. Since the threshold is $Q(n, h) - \tau + R_0$, the equilibrium participation rate satisfies:

$$n = Q(n,h) - \tau + R_0$$

Substituting the definition of $Q(n,h)$ gives:

$$n = b_0 + \mu n - \lambda h - \tau + R_0$$

Therefore:

$$n = \frac{b_0 - \lambda h - \tau + R_0}{1 - \mu}, 0 < \mu < 1 \tag{1}$$

Equation (1) is the participation fixed point, making the institutional tradeoff explicit. Participation is higher when the club is more valuable ex ante, when the community is thicker, and when newcomer status carries some continuation value. Participation is lower when incivility degrades the commons, or when the club's boundary rules impose excessive friction.

The conduct margin follows Friedman and Resnick more directly. Suppose each active user chooses between civil and uncivil behaviour. An uncivil act yields a one-period private gain g_i, which may represent expressive utility, attention, partisan signalling, harassment value, or simply amusement. If the act is detected with probability $p \in [0,1]$, the user loses good standing and must remain either

[3]The model below is not a restatement of Friedman and Resnick's (2001) repeated random-matching game; rather, it is a reduced-form adaptation of their core mechanism – cheap identity reset weakens reputational discipline and pushes communities toward a newcomer-tax equilibrium – embedded in a club-governed discourse environment.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

excluded or re-enter as a newcomer. Let $\kappa \geq 0$ denote the cost of identity reset, re-verification, or recovery after sanction. Let $R > R_0$ denote the continuation value of good standing inside the club. Then, a user chooses uncivil behaviour if and only if:

$$g_i > \delta p(\kappa + R - R_0) \tag{2}$$

where $\delta \in (0,1)$ is the discount factor. The right-hand side is the expected future cost of misconduct. Cheap, disposable identities correspond to low κ. Weak reputational continuity corresponds to a small gap $R - R_0$. Stronger moderation or attribution raises p. Equation (2) is the Friedman-Resnick mechanism expressed to best understand contemporary identity architectures.

Assume now that $g_i \sim U[0, \overline{g}]$. Then, the share of users who choose uncivil behaviour is:

$$h = 1 - \frac{\delta p(\kappa + R - R_0)}{\overline{g}}, 0 < \delta p(\kappa + R - R_0) < \overline{g} \tag{3}$$

Substituting (3) into (1) yields a key equilibrium equation:

$$n = \frac{b_0 - \lambda + \lambda \delta p(\kappa + R - R_0)/\overline{g} - \tau + R_0}{1 - \mu}, 0 < \mu < 1 \tag{4}$$

Participation rises when the expected future penalty for misconduct rises because the equilibrium share of uncivil behaviour falls and the quality of the discourse commons improves.

Next, decompose the continuation value of good standing as:

$$R = \overline{R} + \sigma$$

where $\sigma \geq 0$ is the quasi-rent generated by non-transferable, reputation-bearing club rights. In the present context, σ is the value of maintaining a soulbound token linked standing: access to verified-only spaces, higher trust in moderation, governance eligibility, and so on, that cannot simply be sold or moved to a fresh account. Under that interpretation, we have:

$$\frac{\partial h}{\partial \sigma} = -\frac{\delta p}{\overline{g}}, \; and \; \frac{\partial n}{\partial \sigma} = \frac{\lambda \delta p}{(1 - \mu)\overline{g}} > 0 \tag{5}$$

Equation (5) shows that soulbound tokens matter because they make good standing durable, non-transferable, and therefore sanctionable. Their value lies in the future rights and access they secure. An increase in σ reduces uncivil conduct by raising what a user stands to lose from misbehaviour, and the resulting improvement in discourse quality increases equilibrium participation

The model also clarifies the externality. Suppose each active participant makes one contribution so that there are n contributions total. One additional uncivil contribution raises h by $1/n$. Since aggregate utility from the shared discourse environment is $nQ(n,h)$, the direct effect of one or more uncivil contributions on the community's gross participation value is:

$$\Delta[nQ] = -n\lambda\left(\frac{1}{n}\right) = -\lambda$$

Now, let $m > 0$ denote the moderation, enforcement, or adjudication cost generated by one uncivil contribution. Then, the marginal social damage of one additional uncivil contribution is:

$$D = \lambda + m \tag{6}$$

A socially efficient rule therefore would allow uncivil behaviour only when:

$$g_i > D \tag{7}$$

Under the identity regime, however, the actor compares g_i to the private expected penalty in Equation (2), not to D. The governance technology internalizes the externality if:

$$\delta p(\kappa + R - R_0) \approx \lambda + m \tag{8}$$

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

If equality holds exactly, the private and social thresholds coincide. If the left-hand side is smaller, the regime under-deters. If it is larger, the regime over-deters. Because Equations (1) and (4) imply that incivility also depresses participation by lowering the quality of the environment, $D = \lambda + m$ should be read as a lower bound on the full social harm.

A final comparative static involves holding the equilibrium level of incivility fixed and p constant, then differentiating Equation (3) implies:

$$dh = 0 \rightarrow dR_0 = d\kappa + dR \tag{9}$$

Equation (9) formalizes the 'newcomers paying dues' logic.

When identities are cheap to reset, online communities face a distorted choice: either tolerate more low-value abuse or impose diffuse suspicion on newcomers by withholding trust and standing until they have effectively 'paid dues'. Institutions that raise the cost of identity reset and increase the value of good standing relax that tradeoff. This interpretation is consistent with the empirical patterns discussed earlier: stable and attributable identities tend to temper aggressive behaviour, while blunt real-name mandates often reduce participation by raising privacy and entry costs for all users rather than only for bad actors (Cho and Kwon, 2015; Qian *et al.*, 2024; Santana, 2014). Verified pseudonymity is useful because it raises the expected cost of incivility, ban evasion, and Sybil behaviour while preserving pseudonymous participation. Proof-of-personhood raises κ, durable reputation-bearing credentials raise $R - R_0$, and privacy-preserving identity architecture adds a smaller increase in τ than blanket real-name or centralized verification systems.

Governance credentials and verified pseudonymity

One implication of the model is that the relevant institutional question is not whether platforms should require universal public identification, but how they can attach continuation value to good standing without imposing excessive privacy and participation costs. In property-rights terms, soulbound tokens should not be analysed as ordinary crypto-assets. They are better understood as non-transferable governance credentials. An SBT creates a limited bundle of eligibility rights, signalling rights, and sanctionability conditions. It may confer eligibility to enter a forum, access higher-trust channels, or participate in governance; it may permit the holder to signal verified standing, tenure, or credentials; and it may expose the holder to suspension, downgrade, or exclusion if rules are violated. The complementary authority to issue, interpret, update, and revoke those rights lies elsewhere – with the relevant platform, community, consortium, or credential issuer. Enforcement therefore does not come from the token alone. It comes from the surrounding institutional arrangement: admission rules, moderation procedures, revocation standards, and appeals mechanisms. Here, the economic significance of an SBT is not exchange value, but its ability to attach a non-transferable continuation value to good standing (Ohlhaver *et al.*, 2022).

Comparative institutional context and design requirements

The preceding section framed online discourse as a club-governed informational commons in which the relevant institutional problem is not anonymity in the abstract, but the ease with which present conduct can be separated from future consequences. This section compares governance arrangements by asking how each raises the expected cost of abuse, ban evasion, and inauthentic participation without destroying the participation and privacy benefits that pseudonymous speech can provide. In particular, the goal is to study feasible institutional bundles and weigh the costs and benefits.

Governance alternatives as institutional bundles

The relevant governance alternatives differ along several margins at once: admission costs, privacy costs, monitoring costs, sanction credibility, portability of standing, concentration of control, and the

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

procedural stakes of exclusion. Open or weakly authenticated pseudonymity remains attractive because it keeps participation costs low and preserves separation between public speech and civil identity. But precisely because accounts are easy to create, abandon, and replicate, it also leaves communities with low reset costs, high monitoring burdens, weakly credible sanctions, and recurrent exposure to Sybil behaviour and ban evasion. Blanket real-name mandates attack the reset-cost problem more directly, yet they do so by imposing privacy and retaliation costs on all users rather than selectively on bad actors. Centralized platform verification or KYC softens the publicity problem by allowing users to remain publicly pseudonymous, but it duplicates admission burdens across firms, creates firm-specific stores of sensitive identity data, and expands the discretionary power of platforms over access and exclusion. Purely behavioural or algorithmic moderation can remove, rank, or suppress content without collecting identity at all, and is therefore an important baseline, but without durable identity it disciplines content episode by episode rather than attaching consequences to the actor over time. Decentralized verified pseudonymity is thus best understood as a fifth type of bundle: private authentication, public pseudonymity, more durable standing (from a reputational sense), and a more distributed allocation of control.

Trade-offs are inevitable; no bundle dominates on every margin. Real-name rules and centralized KYC can improve attribution and sanction credibility, but they do so by concentrating both personal data and gatekeeping power. Open pseudonymity preserves entry and protects vulnerable speech, but it makes serial re-entry and one-to-many account multiplication relatively cheap. Algorithmic moderation scales, but it is not a full substitute for attributable standing. As the durability of identity increases, so do the stakes of wrongful exclusion, which makes appeal, recovery, and procedural constraint more rather than less important. The design problem is to raise the expected cost of misconduct without making participation contingent on public identification.

Durable standing and civility

Research on online disinhibition should not be read as a blanket indictment of pseudonymous speech, but as evidence that dissociation from ordinary social consequences can reduce restraint (Suler, 2004). Santana (2014) finds higher incivility where commenting is anonymous; Coe *et al.* (2014) show that regular pseudonymous participants tend to be more civil than infrequent drive-by commenters; Rösner and Krämer (2016) find that the effect of identifiability is mediated by prevailing group norms rather than operating mechanically; and Qian *et al.* (2024) show that when virtual identity becomes more visible, even lightweight engagement becomes more selective. In this sense, the empirical literature does not imply that public legal-name disclosure is necessary for accountability, but rather continuity of standing for reputational considerations matters.

Persistent pseudonyms can perform some of the same disciplining work as real names because they allow others to condition trust, access, and future interaction on past conduct. Repeated interaction does not require public-name disclosure; it requires that conduct be attached to an identity that is costly to discard. What weakens discipline is not pseudonymity as such, but the low cost of abandoning a tarnished persona and returning under a fresh one. The distinction has both normative and practical implications. For whistleblowers, political minorities, dissidents, workers subject to employer sanctions, and users facing harassment or retaliation, pseudonymity has independent value. A governance regime that equates accountability with public naming therefore overshoots. The institutional implication is that the mechanism should preserve public pseudonymity while making continuity of standing and identity reset materially consequential.

Bakshy *et al.* (2015) show that on Facebook – a platform that has already been organized around real-name social graphs – exposure to cross-cutting political content is constrained by homophily and ranking. Bail *et al.* (2018) show that direct exposure to opposing views on Twitter/X can intensify, rather than reduce, polarization. Sunstein and Hastie (2015) point to group polarization, informational cascades, and hidden-profile dynamics that operate even in settings with known identities. Mildenberger (2018) similarly emphasizes how conflict-kindling institutions can sustain antagonistic

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

equilibria. Mildenberger and Tingley (2019) add that second-order beliefs – beliefs about what others believe – can themselves sustain conflict and inaction even when first-order preferences are less extreme. These mechanisms are only partly related to identifiability, rooted more deeply in group dynamics, network structure, and institutional context.

Identity can matter more for the authenticity margin than on the belief-formation margin. Shao *et al.* (2018) show how bots and other inauthentic actors can amplify low-credibility content and distort perceived support. Cheap pseudonyms, account farms, and serial ban evasion can make fringe or abusive conduct appear more representative than it is, thereby altering second-order beliefs about public opinion and community norms. Proof of personhood and durable standing can matter, but mainly by reducing fake amplification, serial harassment, and inauthentic participation. Purely behavioural moderation can demote particular content, yet without durable identity it remains weak against repeat entry and synthetic crowd effects. Verified identity is justified as a tool for civility, accountability, and authenticity, not for reconciling real ideological disagreement.

Constitutional and organizational limits of public identification

South Korea provides an illustrative example because it confronted the online-accountability problem in unusually explicit institutional form. To set the stage, consider that by the mid-2000s, large portals and internet news sites had become central channels for political participation and public commentary, while concerns about rumour, defamation, and cyberbullying were increasingly framed as problems of weakly attributable speech. Korea first moved in the electoral context, where identity-verification rules were applied to election-related sites, and then generalized that logic through the Network Act to major domestic forums. The governing intuition was as follows: if posting required prior verification against a civil identifier, abusive speech would become less attractive ex ante and easier to police ex post.[4] In that sense, the Korean case is valuable not because it tells us whether 'identity' matters in the abstract, but because it shows what a mature attempt to operationalize accountability through front-end verification actually looked like in practice (Constitutional Court of Korea, 2012; Leitner, 2009; Lyou, 2013).

Importantly, the Korean regime was not merely a norm of public-name display. Its institutional core was compulsory prior verification and platform-side retention. Users of covered message boards had to authenticate through resident-registration numbers or comparable credentials, and once the threshold was lowered to sites averaging 100,000 daily users, the number of covered services expanded sharply. In the Court's own summary, the system required service providers not only to verify identity but also to retain verification information for up to six months after a posting disappeared, and potentially longer where the posting remained online. Korea tested a much stronger bundle than ordinary 'real-name' rhetoric suggests, making domestic intermediaries responsible for collecting, storing, and operationalizing civil identity as a condition of participation in large-scale online discourse (Constitutional Court of Korea, 2012; Leitner, 2009).

The evidence cited by later commentators and by the Constitutional Court suggests that the regime dampened participation more clearly than it improved discourse quality: comment activity declined, users had reason to shift toward overseas or uncovered services, and there was no convincing showing that illegal or malicious postings had fallen substantially. The Court accordingly accepted the legitimacy of addressing defamation and related harms, but held that blanket verification was disproportionate because less restrictive tools already existed, including deletion, temporary blocking, post hoc tracing, civil remedies, and criminal punishment. Just as importantly, the Court treated anonymous and pseudonymous expression as constitutionally significant and stressed that mandatory retention of verification data burdened users' control over personal information while increasing the

[4]The 2012 Constitutional Court decision is also analytically important as an institutional check on overreach. The Court accepted the anti-defamation purpose of the law but held that the regime was disproportionate because less restrictive enforcement tools already existed, evidence of efficacy was weak, the law required retention of verification data, and the burden fell unevenly on domestic providers as users migrated to foreign services.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

risks of misuse or leakage. The Korean case therefore does not show that accountability requires public naming, but rather that there are legal and organizational limits of a regime that pursues accountability through centralized retention of civil identity (Constitutional Court of Korea, 2012; Cho and Kwon, 2015; Lyou, 2013).

Seen in that light, centralized platform verification or KYC is less a clean alternative than a narrower variant of the same organizational logic. It is narrower because users can remain publicly pseudonymous, but it still concentrates identity collection, storage, and sanction administration inside private intermediaries – it is still fundamentally a centralized regime. That may improve authentication where bot pressure is severe, yet it also creates firm-specific stores of sensitive data, multiplies verification burdens across platforms, raises switching costs, and makes suspension, recovery, and rehabilitation dependent on the unilateral governance of platform operators. What centralized verification gains in immediate sanction credibility it often loses in portability, privacy, and due process. In this sense, a key lesson is not that private authentication should be rejected, but that authentication should not depend on single-platform custody of civil identity.

Thus, proof-of-personhood, decentralized identifiers, and verifiable credentials are the relevant building blocks for the mechanism developed below. Proof-of-personhood addresses the one-person-many-accounts margin by raising the cost of Sybil behaviour. DIDs provide persistent, user-controlled identifiers that need not disclose a civil name. Verifiable credentials allow an issuer to attest a limited fact – such as uniqueness, age threshold, or good standing – without requiring wholesale disclosure of personal data. Anchored to non-transferable credentials or SBT-style attestations, these tools separate functions that blanket real-name rules and centralized KYC tend to collapse: uniqueness without public naming, portability without total platform dependence, and continuity of standing without freely transferable reputation assets (Borge *et al.*, 2017; Ohlhaver *et al.*, 2022; W3C, 2022, 2025). They do not eliminate coercion, credential rental, moderation error, or political abuse – which remain governance problems – but they enable a better institutional arrangement: private authentication paired with public pseudonymity and durable standing.

Legal scope, data governance, and design requirements

This paper focuses on constitutional democracies, such as the United States, the European Union, and similar rule-of-law settings, in which pseudonymous participation retains independent legal or constitutional value and exclusion cannot be treated as a purely technical event.[5] In more authoritarian or protectionist settings, digital identity can be folded into censorship, domestic gatekeeping, and political surveillance. Those cases remain important as a warning, but they pose a different institutional problem from the one examined here (Yalcintas and Alizadeh, 2020). Thus, I explore a rights-constrained governance in which public pseudonymity retains both expressive and instrumental value – not a state-administered universal identification.

In this sense, the legal question is not whether 'blockchain' is lawful in the abstract, but rather whether a particular identity architecture allocates obligations, limits data capture, and preserves contestability in a form that can survive ordinary legal review. In the European context, that legal frame is distributed across several domains rather than contained in a single body of law. General Data Protection Regulation (GDPR) is most relevant to personal-data processing, data minimization, and controller allocation; the DSA to platform governance, procedural duties, and systemic-risk management; European Digital Identity Regulation (eIDAS 2.0) to digital identity and attestations; and Markets in Crypto-Assets Regulation only where transferable tokenized community-currency mechanisms are used. The purpose here is not to provide a full doctrinal treatment of each regime, but

[5]The argument is not intended to apply to centralized state identity regimes in which digital identity is directly integrated with censorship, political surveillance, or comprehensive social control. In those settings, the institutional problem is different. The question is no longer how to preserve accountability without public naming, but how identity infrastructure itself can become an instrument of coercion.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

to clarify that the proposed architecture raises distinct legal questions that should not be collapsed into a single inquiry about whether blockchain-based identity is lawful.

Finck (2019) argues that distributed ledgers are difficult to reconcile with familiar GDPR categories (e.g., erasure and data minimization because responsibility is diffused and governance is often underspecified. Giraudo's (2022) account of legal bubbles sharpens the broader institutional point. Systems built on provisional assumptions about the lawful collection, correlation, and quasi-propertization of personal data may rest on unstable legal foundations that later courts, regulators, or constitutional review narrow or reverse. Alston *et al.* (2022) add a complementary lesson for institutional design: blockchain systems are not self-executing substitutes for governance, but evolving polycentric orders subject to internal rule change and external legal constraint. An identity regime that presumes stable entitlement to store, correlate, or indefinitely exploit personal data therefore rests on uncertain legal foundations.[6]

The prudent design response is accordingly modest. Personal data and linkable credential material should remain off-chain or otherwise outside publicly replicated ledgers. Nor is mere hashing a complete escape hatch, since whether a reference is genuinely anonymous or merely pseudonymized depends on the possibility of relinking it to a person within the broader architecture. Where on-chain references are used, they should serve only as non-identifying status markers or pointers whose practical significance depends on separately governed off-chain records. Controller responsibility should likewise be allocated as clearly as possible: credential issuers should bear responsibility for the personal data they verify and retain, while relying platforms should bear responsibility for the additional data they collect and the sanctions they impose. Public-facing credentials should reveal no more than is necessary for the immediate governance function and, where possible, should be presented through selective disclosure or zero-knowledge proofs rather than as data-bearing tokens. Because the GDPR is technology-neutral but not self-applying, co-regulatory devices such as codes of conduct and certification mechanisms may also matter as complements to architectural restraint (Finck, 2019).

Because exclusion errors are unavoidable, any sanction regime must include notice, reason-giving, appeal, and some path to rehabilitation rather than irreversible blacklisting. Revocation should ordinarily operate by rendering a credential invalid or non-presentable, not by pretending that an immutable public history can simply disappear. Keys and accounts should be recoverable, and key rotation should be possible without destroying standing. Interoperability should allow users to carry proofs of standing across willing communities, but it should not create automatic universal cross-platform surveillance or a single pan-platform sanctions file. The requirements for the mechanism that follows are as follows: it should preserve public pseudonymity rather than require public real names; it should raise the cost of Sybil attacks, ban evasion, and serial harassment by making standing durable and reset costly; it should minimize on-chain personal data; and, it should remain contestable and procedurally constrained rather than treating identity governance as self-justifying once technically implemented (Alston *et al.*, 2022; Finck, 2019).

Mechanism design: verified pseudonymity as a hybrid governance architecture

The proposed identity architecture is a governance arrangement that reallocates verification, admission, attribution, exclusion, record-keeping, and appeal across users, credential issuers, platforms, and governance bodies – not a technical supplement to moderation. That is the relevant sense in which blockchain-based identity is an institutional technology rather than a stand-alone piece of infrastructure (Davidson *et al.*, 2018). It changes who can make credible claims about personhood and

[6]Alston et al. (2021) provide a more granular decomposition, distinguishing four interrelated layers of blockchain governance: protocol governance, subsidiary governance, competitive governance, and 'superior' governance in the form of the surrounding legal and regulatory order. That framework shows that even permissionless systems remain nested within broader institutional constraints rather than operating outside governance and law.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

standing, who may rely on those claims, and under what conditions sanctions can be imposed and reviewed. As Harris (2018) and Safner (2016) suggest in related work on digital commons, durable cooperation depends on boundary rules, information rules, and sanction rules, not on protocol design alone. For that reason, the architecture is better described as a polycentric order than as a fully decentralized one (Alston *et al.*, 2022). It is also necessarily hybrid: a workable system combines code with organizational discretion, private or public issuers with platform rule-making, and automated checks with human review (Frolov, 2021). Users control when they present credentials. Issuers verify underlying facts and maintain the associated records. Platforms decide which credentials matter for access, ranking, and moderation. Governance bodies, consortia, or appeals panels constrain arbitrary exclusion and define the terms on which credentials can be recognized across communities.

This section operationalizes the logic developed earlier on the economics of incivility and accountability where proof-of-personhood raises the cost of identity reset by making serial re-entry and one-person-many-accounts strategies more expensive (see Figure 1). Decentralized identifiers and selective disclosure contain onboarding and participation costs because they allow authentication without requiring each platform to collect and store civil identity. Reputation-bearing, non-transferable credentials increase the continuation value of good standing because verified status becomes durable and not cheaply alienable. Moderation institutions, issuer governance, revocation rules, appeals, and account recovery affect by changing how credibly misconduct can be detected, attributed, and sanctioned. The question is therefore whether this bundle reduces transaction costs, economizes on information, and strengthens sanction credibility more effectively than the alternatives of open pseudonymity, blanket real-name rules, centralized KYC, or purely behavioural moderation?

The label 'SBT' is retained because it is already used in the paper, but analytically the relevant object is a governance credential. A soulbound-token-style credential is best understood as a bundle of eligibility rights, signalling rights, and sanctionability conditions. It may make a user eligible to enter a verified forum, use higher-reach features, or signal tenure, expertise, or good standing. It also places that user within a regime under which misconduct can lead to suspension, downgrade, loss of privileges, or a requirement to requalify. Crucially, however, the token alone does not guarantee enforcement – a common misunderstanding about blockchain, as pointed out by Alston *et al.* (2022). The relevant authority lies in the surrounding institutions: the issuer that verifies the claim, the platform that interprets it for a particular purpose, and the governance body that reviews error or abuse. The mechanism is therefore not 'on-chain enforcement', but rather institutional enforcement using a persistent credential layer.

Proof-of-personhood and the cost of identity reset

Proof-of-personhood refers here to any arrangement that allows a relying party to verify that an account is backed by a distinct human rather than by a bot, a script, or a duplicate identity. The governance problem it addresses is basic. If an excluded user can re-enter immediately under a new handle, or if one actor can cheaply multiply accounts, exclusion loses force and reputational discipline weakens. In the model on the economics of incivility and accountability, proof-of-personhood works mainly by increasing the cost of reset. That increase depends on institutional choices about what exactly must be unique. A community may want one human per governance vote, one human per high-reach account, one human per moderation role, or simply a sufficiently costly process for obtaining additional identities. Proof-of-personhood is a rule about entry and rate-limiting – not a solution to trust.

That service can be supplied through different institutional forms (Alston *et al.*, 2022; Frolov, 2021). A public issuer may verify uniqueness against civil records. A private provider may run a one-time verification service and issue a reusable credential. A community association may rely on attestors who know local participants and can vouch that an applicant is not a duplicate. A consortium may accredit multiple issuers and let platforms accept credentials from any issuer that satisfies common standards. Each approach shifts costs and risks differently. Public issuers may broaden coverage but increase

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

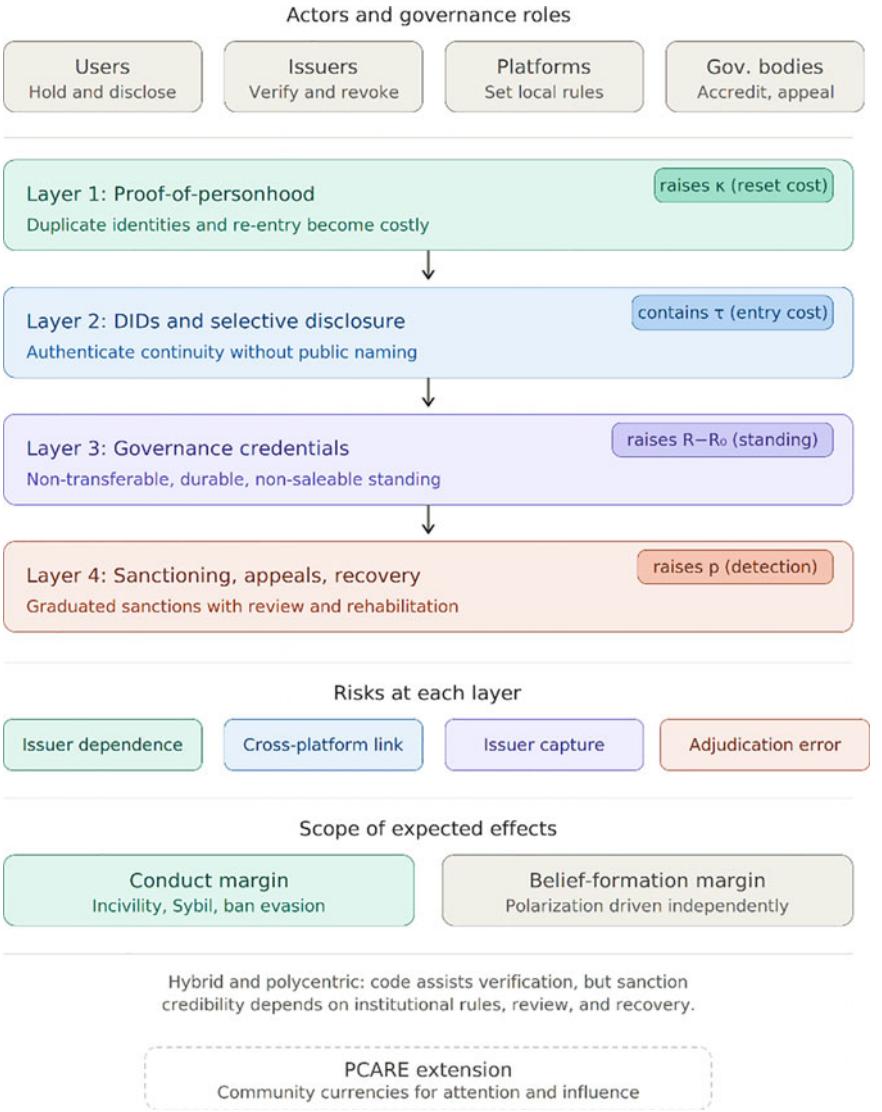


Figure 1. Verified pseudonymity as a hybrid governance architecture. *Notes*: Each layer maps to a parameter in the model on the economics of incivility and accountability. The architecture reallocates governance functions across users, credential issuers, platforms, and governance bodies. Colour-matched risk pills identify the principal failure mode at each layer. The bottom panel distinguishes the conduct margin, where the mechanism has a comparative advantage, from the belief-formation margin, which is driven independently by homophily, selective exposure, and informational cascades. Source: Author's production.

dependence on the state. Private issuers may scale more easily but create commercial gatekeepers. Community attestors may use local knowledge effectively but can reproduce favouritism or exclusion. Consortia can reduce dependence on a single provider, but they introduce coordination and accreditation costs. The institutional question is therefore who has authority to attest uniqueness, what evidence counts, and how errors are reviewed.

For relying platforms, the information obtained from this layer should be limited. In many cases, the platform needs to know only that the account satisfies a uniqueness threshold or rate-limit condition. It does not need the person's legal name, scanned documents, or biometric record. That separation matters because it can reduce moderator screening costs and exposure to bot swarms without turning

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

each platform into a repository of civil identity. It also matters for participation. A high-friction personhood process may raise the cost of identity reset for bad actors, but it can also raise the onboarding and participation costs for ordinary users. The design problem is to make duplication and re-entry costly without making honest entry unnecessarily difficult.

Decentralized identifiers (DIDs), selective disclosure, and the separation of verification from publicity

Once uniqueness is addressed, the next problem is to authenticate continuity without requiring public naming. Real-name rules and many KYC systems collapse three distinct questions into one: whether the participant is unique, whether conduct is attributable across time, and whether outsiders can see who the participant is. Verified pseudonymity separates those questions. A decentralized identifier, in the W3C sense, is a persistent identifier under user control rather than platform control (W3C, 2022). In institutional terms, it allows a platform to authenticate a stable pseudonymous participant without having to become the custodian of that person's civil identity.

The governance value of that separation is clearest if uniqueness, attribution, and publicity are kept analytically distinct. Uniqueness concerns whether additional identities are sufficiently costly to deter Sybil behaviour. Attribution concerns whether today's conduct can be linked to yesterday's conduct for purposes of warning, reward, or sanction. Publicity concerns what the public learns about the person behind the account. Proof-of-personhood addresses the first problem. DIDs address the second. Selective disclosure limits the third. A platform can therefore know that it is dealing with the same credential-bearing participant over time without learning the participant's legal name and without maintaining its own civil-identity database.

This is also the layer that helps contain onboarding and participation cost relative to blanket real-name rules or platform-specific KYC. The W3C verifiable-credentials model divides roles among issuer, holder, and verifier (W3C, 2025). The issuer verifies a fact. The user holds the credential. The platform verifies only the claim it needs for the immediate governance purpose. A forum that needs to know whether a user is over eighteen, is a unique person, or is in good standing does not need the underlying dossier from which that fact was derived. That reduces duplicative onboarding, lowers data-handling and compliance costs for platforms, and allows users to reuse attestations across settings. Verification can occur without requiring each platform to store more personal data than its local governance task requires.

Verifiable credentials and SBTs as non-transferable standing

A verifiable credential is an issuer-signed attestation about some fact or status of the holder. It may certify that the holder is above a certain age, has completed an identity check, belongs to a profession, or has remained in good standing within a particular community. An SBT-style credential adds a further design feature: it is non-transferable, i.e., it cannot be sold or freely moved as an ordinary asset (Ohlhaver *et al.*, 2022). This should not be overlooked since key economic object is not information by itself, but the presence of durable standing. Stable pseudonyms already allow repeated interaction. Non-transferable credentials make the future value of that interaction more difficult to detach from the history that produced it.

This is where the mechanism most directly increases the continuation value of good standing, $R - R_0$. Formally, write the continuation value of participation as $R = \overline{R} + \sigma$ where σ is the quasi-rent attached to verified standing. That quasi-rent may consist of eligibility to post in higher-trust spaces, access to high-reach features, lower screening frictions, governance participation, or simply the practical advantages of a long-tenured account in good standing. Because the credential is non-transferable, this quasi-rent cannot be sold, moved to a fresh account, or detached from the conduct history that generated it. A user who has accumulated such standing has more to lose from misconduct than a user operating through a disposable handle. That is the relevant contribution of SBT-style

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

design. It is not that tokens somehow cause better behaviour. It is that non-transferable standing makes the future consequences of present conduct more durable.

The particular claims embodied in credentials can vary. Communities may rely on credentials for personhood, age thresholds, one-time identity verification, expertise, tenure, or local good standing. The economic role of those credentials is not the metadata alone. It is to reduce information costs about traits the community cares about and, more importantly, to attach consequences to the actor rather than only to an isolated content episode. That is also why the contribution of SBTs is more specific than the general value of pseudonymity. Pseudonymity preserves privacy and repeated interaction. Non-transferable standing strengthens the intertemporal incentive to preserve one's position within the community.

Recognition of such standing does not need to achieve universality to be effective; a marketplace can and should take form. For example, one platform may treat a tenure credential or expertise credential as relevant; another may not. Issuers verify claims and remain responsible for updating or revoking them. Users decide when to present them. Platforms decide which ones to honour and for what purpose. Governance bodies may accredit issuers or define common credential classes. This limited and plural recognition is important institutionally. It preserves room for local rule-making and avoids collapsing all standing into a single universal rating.

Platform integration, sanctioning, appeals, and recovery

The final layer makes the previous layers consequential. Platforms remain the first-line governors of their own clubs. They decide whether proof-of-personhood is required for posting, whether stronger credentials are needed for higher-reach functions, and which forms of conduct trigger warnings, throttles, suspension, or exclusion. Issuers do something different. They verify underlying facts, maintain the evidentiary basis for the credentials they issue, and revoke or update those credentials when the claim itself has become false, compromised, or fraudulently obtained. Governance bodies or consortia perform a third function: they define interoperability rules, accredit issuers, hear appeals, and limit arbitrary cross-platform enforcement. The architecture is governable only if those functions remain distinct.

At the point of authentication and access control, the arrangement can reduce transaction costs and data concentration. A platform can require proof of personhood for ordinary posting, an age credential for access to regulated features, or a stronger identity credential for high-reach functions without collecting the underlying civil records itself. Because credentials are reusable, users do not need to repeat the entire verification process at every site. Because the platform verifies only a bounded claim, it does not need to warehouse as much personal data as a centralized KYC regime. The arrangement can also lower information costs for moderation. A newcomer backed by a valid personhood credential is not identical to an unverified drive-by account, and a user in good standing need not be screened in the same way as a newly created disposable account.

The effect on attribution, p, is more important. Misconduct becomes more costly only if detection, attribution, and sanction are credible. Durable identity raises p almost by construction because conduct can be linked across incidents and sanctions can target the participant rather than a disposable handle. But that increase is not automatic, because moderation remains an incomplete-contracting problem. Howell and Potgieter (2021) are useful here. Whether a communication is satire, threat, harassment, quotation, or evidence often depends on context that cannot be fully specified in advance. Coordinated abuse may be dispersed across many small acts. Rehabilitation requires judgment about changed conduct. Automated systems can help with authentication, rate limits, and some clear rules. They cannot fully replace adjudication.

Decentralized moderation juries or similar bodies should be analysed as organizational forms rather than as technical features. Davidson (2025) also argues that governance arrangements must solve ordinary organizational problems: who selects decision-makers, what evidence they see, how conflicts of interest are handled, what standards of review apply, and how decisions can be revised. The same questions arise here. A platform review team, an industry consortium panel, an independent ombuds

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

function, and a token-selected jury are different ways of organizing review, not interchangeable modules. Their costs, legitimacy, and exposure to capture differ.

A workable sanctions ladder should therefore be graduated, claim-specific, and reviewable. Minor violations may justify warnings, reduced reach, temporary throttling, or temporary loss of features. Repeated or severe violations may justify suspension of good-standing credentials or exclusion from a particular forum. Platform sanctions and issuer sanctions should not be conflated. A platform may exclude a user from one community without invalidating the user's underlying personhood credential. Issuer revocation should usually be reserved for compromised credentials, false claims, or narrowly specified cases in which the claim itself has become unreliable for relying parties. Cross-platform sanctions should be the exception, tied to a clearly defined claim, and subject to appeal. A finding of repeated violent threats may justify temporary exclusion from certain high-risk features across cooperating platforms. A dispute over tone, ideology, or ordinary moderation judgment should not become a portable blacklist. As identity durability increases, wrongful exclusion becomes more costly. Due process becomes more important, not less.

The same logic applies to rehabilitation and technical recovery. A regime that makes exit costlier must also provide a route back from error, both moral and technical. Users need notice, reasons, and an opportunity to contest evidence. Many other design features also remain. For example, some sanctions should expire; some credentials should be restorable after a period of compliance; some findings should remain local to the community in which they arose. Account recovery is equally central. If a user loses control of a device or key, the answer cannot be automatic forfeiture of years of standing. Yet recovery cannot be so permissive that it becomes a cheap reset channel. Social recovery, in which a pre-designated set of trusted parties can help restore account control, and issuer-assisted reissuance are possible responses, but both require clear procedures and audit trails. This is another respect in which the arrangement is hybrid: technical continuity has to be backed by institutional procedures for review, recovery, and redress.

Crucially, a decentralized credential ecology is not the same thing as a centralized social credit system. Devereaux and Peng (2020) are useful on this point. The distinction is institutional. A decentralized ecology has plural issuers rather than a single sovereign rater, limited-purpose credentials rather than one universal file, contestability among relying communities, and no single database that aggregates all conduct into a common score. That distinction can erode if one issuer becomes dominant, if credentials become universally correlated, or if cross-platform sanctions become routine. The difference therefore has to be built into the governance structure.

Still, decentralization need not be inherently welfare-improving. Hybrid governance brings coordination costs, standards fragmentation, accreditation disputes, and the risk that nominally decentralized systems recentralize around dominant issuers, wallet providers, or platforms. Wallet recovery and key management remain practical obstacles. Some users will lack acceptable documents, devices, or access to approved issuers. Portability can also produce overexclusion if sanctions travel farther than the underlying misconduct warrants. Verified pseudonymity may be more likely in liberal democracies to outperform open pseudonymity, blanket real-name rules, centralized KYC, or purely behavioural moderation on some margins and in some settings. It is not a general synonym for better governance. Once participation and standing are organized through verified pseudonymity, a further question arises inside the verified community: how scarce attention, bandwidth, and governance rights should be allocated among participants who have already crossed the identity threshold. The next section introduces a fifth layer to address that problem. It asks how priced access and governance rights can be separated within a verified community, building on the architecture developed here rather than replacing it.

Plural community asset resource exchange (PCARE)

The four-layer architecture in Section mechanism design: verified pseudonymity as a hybrid governance architecture solves only part of the governance problem. Proof-of-personhood, decentralized

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

identifiers, verifiable credentials, and platform integration can authenticate participants and make sanctions durable across time, but they do not by themselves allocate scarce attention, price congestion, or prevent verified insiders from dominating governance. A community of fully verified users can still be overwhelmed by low-value posting, coordinated amplification, frivolous appeals, or agentic swarms. The remaining problem is institutional rather than merely technical: how should a club-governed informational commons ration scarce bandwidth and decision rights once admission has occurred? Ohlhaver's (2025) PCARE framework is an operational extension that helps separate two margins that digital platforms often collapse into one – access to attention and control over the rules.

Why identity alone may be insufficient

Verified pseudonymity raises the cost of identity reset and increases the continuation value of good standing. That solves part of the incivility problem, but not the allocation problem. Moderator time, front-page placement, high-reach channels, API throughput, and appeal capacity are scarce. When these resources are unpriced, they are rationed by unmanaged congestion, informal status competition, or centralized platform discretion. Nor does one-person-one-credential solve the capture problem. A verified participant can still flood a forum, repeatedly purchase distribution, or coordinate with others to dominate governance. The unresolved issue is therefore how to separate ordinary participation from scarce access, and scarce access from durable control.

This is precisely where the distinction between attention and influence matters. Attention concerns the right to consume scarce community bandwidth: posting into high-reach channels, initiating large-scale distribution, escalating disputes, invoking human review, or directing autonomous agents to use shared infrastructure. Influence concerns the right to alter the rules governing those activities: setting prices, changing moderation standards, selecting reviewers, or determining sanctions and reinstatement. Market institutions tend to price attention relatively well but can allow money to buy control. Hierarchical institutions can assign control, but often do so opaquely and without credible checks. The attraction of PCARE is that it decomposes these margins rather than allowing one to overshadow the other.

Layer 5: separating the price of attention from the cost of influence

I add a fifth layer to the architecture: a community-currency system that prices attention separately from influence. Let each community c recognize two linked but distinct instruments. The first is a transferable attention token, a_c, used for ordinary resource exchange inside the community and, more specifically here, for acquiring scarce bandwidth. The second is a non-transferable influence stake, s_c, obtainable only by locking attention tokens into a governance position attached to a verified pseudonymous identity. The two instruments share a unit of account, but moving from attention to influence requires foregoing liquidity and accepting exposure to sanction. Influence is therefore not simply bought; it is carried.

The key separation is between scarce attention and durable influence. Proof of personhood and DIDs do not create that separation by themselves. Instead, they establish authorization and provenance: the system can verify that a request comes from a credentialed participant, acting through a persistent pseudonymous identity, without requiring public disclosure of civil identity. PCARE adds the allocation layer. It distinguishes the price paid to consume scarce community bandwidth from the stake required to exercise governance power over the rules that allocate that bandwidth.

Let each community c recognize two linked but distinct instruments. The first is a transferable attention token, a_c, used to acquire scarce bandwidth inside the community: posting into high reach channels, escalating disputes, requesting human review, or directing software agents to use shared infrastructure. The second is a nontransferable influence stake, s_c, obtained only by locking value into a governance position attached to a verified pseudonymous identity. Moving from attention to influence therefore requires giving up liquidity and accepting exposure to sanction. Influence is not simply

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

bought. It is carried within a community specific position that can be downgraded, slashed, or lost if governance power is abused.

This separation has two comparative institutional advantages. First, scarce attention can be priced rather than rationed entirely by platform discretion, informal status, or abusive congestion. Second, liquid wealth cannot convert immediately into governance control. A wealthy actor may still buy attention at the margin, but durable influence requires locking value into a nontransferable, slashable, community specific position tied to an enduring identity. The result differs from ordinary token voting, where liquid capital can often be turned into control with little commitment to the community's long run quality.

Operationally, each request r submitted to community c can be represented as a bundle $(\kappa_i, d_i, p_r, s_i)$. Here, κ_i denotes the relevant credential proof – proof of personhood and any additional attributes required for the action; d_i denotes the DID under which the request is made; p_r denotes the attention payment or commitment attached to the request; and s_i denotes the influence stake, if any, associated with the actor. The first two elements determine authorization and provenance. The third prices access to scarce attention. The fourth determines whether the actor may participate in rule-making and what they stand to lose if that power is abused.

This formulation makes clear what PCARE adds to verified pseudonymity. PoP and DIDs establish who is authorized to act and under what persistent pseudonymous identity. The attention payment prices the scarce resource consumed by the request. The influence stake determines whether the actor has governance power and what sanctionable position backs that power. The architecture therefore separates authorization, bandwidth allocation, and governance authority rather than collapsing them into a single verified account or a single transferable token.

The life cycle of a request

Consider a user – or even a user-directed software agent (increasingly likely with agentic AI) – who wants to post, amplify content, appeal a moderation decision, initiate a group channel, or invoke a human review lane. The first step is authentication. The platform or community verifies the minimum necessary credentials through selective disclosure: proof of personhood for ordinary participation, perhaps an age threshold for certain content, or an institutional credential for higher-trust functions. Publicly, the user remains visible only through a pseudonymous DID.

The second step is lane selection. Basic participation should remain 'free' to avoid pricing ordinary speech. Higher-risk or more resource-intensive actions should carry an explicit attention price schedule $p_c(x,\omega)$, where x is the type of action and ω captures current congestion, expected moderation burden, and risk. A normal reply in a lightly used thread may have zero attention cost. By contrast, mass messaging, rapid posting bursts, trend amplification, live-streaming, escalation to human review, or large-scale API calls by autonomous agents should require higher spend. Communities can preserve inclusiveness by combining this price schedule with baseline allowances, periodic rebates, or modest community income for users in good standing.

The third step is routing. Requests are not ordered by spend alone. They are routed by the interaction of spend and standing. A user with a history of compliance, constructive participation, and no unresolved strikes may obtain cheaper access to higher-trust lanes because their expected governance cost is lower. A user with recent abuse, unresolved sanctions, or suspicious activity may face tighter rate limits, higher prices, or pre-publication review. Identity provides continuity, and the currency layer provides rationing; neither function substitutes entirely for the other.

The fourth step is settlement. Attention payments can be burned, transferred to a community treasury, used to fund moderators and appeals, or partially refunded when a request is resolved without consuming the scarce resource for which it was priced. Appeals are a useful example. A community can require a small refundable attention deposit to invoke human review, returning it when the appeal succeeds. That discourages frivolous filings without foreclosing due process.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

Governance voice, stake, and exit

Governance rights should not follow automatically from holding attention tokens. Instead, a verified participant may lock tokens into an influence stake attached to the underlying personhood root recognized by the community. Multiple public DIDs may still exist for privacy and role separation, but governance weight should aggregate to the underlying verified root through privacy-preserving proofs. Otherwise, DID plurality would simply recreate Sybil behaviour at the governance layer and replicate the initial challenge from another vantage point.

Let governance voice be an increasing but concave function of influence stake, such as $v(s) = \sqrt{s}$. The principle is that concentration of control should become progressively more costly at the margin. Square-root weighting does not eliminate asymmetry, yet it weakens the linear conversion of money into power and is therefore better suited to a club-governed informational commons than one-token-one-vote. Unstaking should also be subject to a meaningful delay or cooldown period. If actors can acquire influence, change rules, and exit immediately, then the system has failed to put a true cost on influence. A positive cost of exit is therefore as important as the price of entry.

A further operational point is that influence should be local by default. Stake in community A should not automatically confer voting authority in community B, even if the same personhood credential underlies both. What may be portable across communities is not raw control, but proofs of standing: for example, that an identity has maintained good standing elsewhere for a certain duration, or that it holds a recognized expertise credential. This preserves interoperability without collapsing the architecture into a single pan-platform hierarchy. It also helps avoid the dystopian possibility that one reputation file follows a person everywhere regardless of context.

Reputation, sanctions, and recovery

Operationally, reputation should not be treated as a homogeneous and unidimensional object. Instead, it is better understood as a limited, multidimensional profile of standing attached to a verified pseudonym, including at minimum tenure in the community, unresolved strikes, current rate-limit status, governance eligibility, moderation service, and any context-specific expertise or access credentials. Communities may weigh these dimensions differently, but disaggregation has two advantages. It makes sanctions more proportionate because they can target the specific margin on which misconduct occurred, and it reduces the danger that verified pseudonymity degenerates into a generalized social-credit file (Devereaux and Peng, 2020).

Sanctions could be graduated. Low-level spam or congestion abuse may justify higher attention prices, reduced rate limits, or temporary exclusion from high-bandwidth lanes. Repeated harassment, brigading, and deceptive amplification may justify strikes, partial loss of governance privileges, or slashing of influence stake. Severe misconduct – credible threats, organized impersonation, or repeated ban evasion – may warrant suspension or exclusion from the affected community. In each case, the sanction should attach to the margin on which the violation occurred: access, bandwidth, governance rights, or standing. Clearly, not every violation requires identity-level exclusion; otherwise, a system has no resilience and flexibility to accommodate shocks.

Because contract incompleteness and contextual ambiguity are unavoidable, severe enforcement cannot be fully automated (Howell and Potgieter, 2021). Notice, explanation, and review are thus essential. A useful distinction arises between community recovery and appeals. Community recovery, in the SBT literature, is best understood as a mechanism for restoring account continuity when keys are lost, compromised, or socially contested (Ohlhaver *et al.*, 2022). Appeals and reinstatement are a distinct adjudicative function. They should be handled by human or hybrid juries drawn from verified participants with sufficient standing, selected in ways that reduce clique capture and conflicts of interest. Durability of identity increases the importance of due process because the cost of mistaken exclusion is higher when future standing matters.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

Delegated action and agentic AI

The value of this extension becomes even clearer in an agentic environment. As software agents become cheap to deploy, the relevant risk is not only one human using many accounts, but one human or organization directing large numbers of agents across multiple communities. A simple verified-user badge is insufficient in that setting. Communities need to know on whose authority an agent acts, how much bandwidth it may consume, and what consequences follow if it violates rules. Under the architecture proposed here, a human or organizational principal holds the root credentials and any community-specific influence stake. The principal can delegate a limited operating credential to a software agent tied to a particular DID, budget, scope, and expiration period. The agent can then spend attention tokens up to that budget to post, query, or invoke services inside the community. But the provenance of those actions remains tied to a verified principal whose standing and, where relevant, influence stake can be affected by serious misuse. While not a panacea, it makes swarming and deniable automation materially more expensive.

The same mechanism also supports polycentric composition. Different communities can set different attention prices, credential requirements, appeal rules, and stake thresholds while still recognizing a common grammar of proofs and payments. Instead of one global visibility market governed by a single platform ranking system, there are many local markets for attention and many local constitutions for influence. That arrangement better fits the Ostromian logic of locally adapted rule systems, and it preserves the Hayekian insight that relevant knowledge about moderation, trust, and context is best utilized when it remains dispersed.

None of this implies that decentralization is costless or always superior. The PCARE extension is a novel governance method for communities and requires subsidiarty in order to avoid possible inequities in who can afford scarce attention, collusion, credential rental, and cartel formation among large stakers. Some participants will preference access to global channels and by extension global currencies that risk undermining the required subsidiarity. Relative to open pseudonymity, identity plus PCARE raises the marginal cost of spam, brigading, and governance capture. Relative to blanket real-name rules or centralized KYC, it does so while preserving public pseudonymity and reducing the need for platforms to store civil identity data. Relative to purely behavioral moderation, it attaches consequences to actors across time rather than to isolated content events.

Conclusion

This paper developed an institutional-economics framework for understanding online incivility as an externality problem rooted in cheap, disposable identities and proposed a governance architecture designed to raise the cost of misconduct without requiring universal public identification. The stylized model shows that when identities are costless to reset, communities face a distorted equilibrium: they must either tolerate more abuse or impose diffuse suspicion on all newcomers. Verified pseudonymity – operationalized through proof-of-personhood, decentralized identifiers, non-transferable governance credentials, and graduated sanctioning with appeals – relaxes that tradeoff by raising the cost of identity reset, containing onboarding friction, increasing the continuation value of good standing, and making detection and attribution more credible. The comparative-institutional analysis suggests this bundle can outperform open pseudonymity, blanket real-name mandates, centralized KYC, and purely algorithmic moderation on the civility, accountability, and authenticity margins. But it should not be expected to resolve ideological polarization, which is independently driven by other factors, like homophily.

Thus, the difference between a governance architecture that strengthens accountability and one that enables surveillance or suppression is institutional, not technological. The same credential infrastructure that deters ban evasion can, if captured by a dominant issuer or repurposed by a coercive state, become a tool for exclusion and control. Decentralized credentialing – plural issuers, limited-purpose credentials, contestable recognition, and no single aggregated score – is not the same as a

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

centralized social credit system, but that distinction must be actively maintained through governance design, not assumed as a property of the technology. There are nonetheless adversarial risks: users may find ways to obtain multiple verified identities, credential issuers may accumulate excessive gatekeeping power, and cross-platform sanctions may travel further than the underlying misconduct warrants. Whistleblowers, dissidents, and vulnerable populations who depend on anonymity face particular risks from any system that raises the cost of pseudonymous participation. However, the relevant question is not whether these risks exist – they inevitably will – but whether the proposed architecture manages them more effectively than the alternatives.

Future research should move from comparative-institutional reasoning to direct empirical evaluation. One approach would be field experiments or phased platform rollouts that compare verified pseudonymity with open pseudonymity, real-name rules, or stronger centralized verification on outcomes such as harassment rates, repeat sanctions, ban evasion, successful appeals, moderator time per unit of content, and participation by new and vulnerable users. A second approach would examine the portability margins more directly by testing whether non-transferable standing reduces re-entry after sanction and whether it improves trust and contribution quality without producing excessive chilling effects. A third agenda would study the PCARE extension by analysing whether separating priced access from governance rights reduces spam, coordinated amplification, and frivolous appeals while limiting governance capture by high-resource actors. Especially in agentic AI environments, the key empirical question is whether coupling identity, tokenized access, and stake-based accountability improves discourse quality and institutional resilience relative to the available alternatives. Besides natural experiments arising from international policies, Decentralized Autonomous Organizations may be an especially fruitful environment to empirically test these ideas.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1744137426100587.

Acknowledgements. Thank you to Joshua Ammons, Glen Weyl, and Puja Ohlhaver for thoughtful comments. The core insight of this paper comes from Ohlhaver *et al.* (2022). I am grateful to the Institute for Humane Studies for their support (grant no. IHS019789). The paper was previously circulated as 'Soulbound Tokens, Identity, and Depolarizing Social Media'. An online appendix for the paper is located at: https://papers.ssrn.com/abstract = 5525618.

References

Alston E., Law W., Murtazashvili I. and Weiss M. (2021). Can permissionless blockchains avoid governance and the law? *Notre Dame Journal on Emerging Technologies* **2** (1), 1–32.

Alston E., Law W., Murtazashvili I. and Weiss M. (2022). Blockchain networks as constitutional and competitive polycentric orders. *Journal of Institutional Economics* **18** (5), 707–723.

Bail C.A., Argyle L.P., Brown T.W., Bumpus J.P., Chen H., Hunzaker M.B.F., Lee J., Mann M., Merhout F. and Volfovsky A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* **115** (37), 9216–9221.

Bakshy E., Messing S. and Adamic L.A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science* **348** (6239), 1130–1132.

Borge M., Kokoris-Kogias E., Jovanovic P., Gasser L., Gailly N. and Ford B. (2017). Proof-of-Personhood: Redemocratizing Permissionless Cryptocurrencies. In IEEE European Symposium on Security and Privacy Workshops, pp. 23–26.

Buchanan J.M. (1965). An economic theory of clubs. *Economica* **32** (125), 1–14.

Celuch M., Oksa R., Ellonen N. and Oksanen A. (2024). Self-censorship among online harassment targets: The role of support at work, harassment characteristics, and the target's public visibility. *Information, Communication and Society* **27** (11), 2171–2190.

Cho D. and Kwon K.H. (2015). The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior* **51**, 363–372.

Coe K., Kenski K. and Rains S.A. (2014). Online and Uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication* **64** (4), 658–679.

Constitutional Court of Korea (2012). Case on the real-name verification requirement on the internet. https://cyrilla.org/api/files/1727687134267j43kge63lj.pdf (accessed 1 March 2026).

Davidson S. (2025). The nature of the decentralised autonomous organisation. *Journal of Institutional Economics* **21**, e5.

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press

Davidson S., De Filippi P. and Potts J. (2018). Blockchains and the economic institutions of capitalism. *Journal of Institutional Economics* **14** (4), 639–658.
Devereaux A. and Peng L. (2020). Give us a little social credit: To design or to discover personal ratings in the era of Big Data. *Journal of Institutional Economics* **16** (3), 369–387.
Finck M. (2019). Blockchain and the General Data Protection Regulation: Can distributed ledgers be squared with European data protection law? European Parliamentary Research Service, Scientific Foresight Unit (STOA).
Friedman E.J. and Resnick P. (2001). The social cost of cheap pseudonyms. *Journal of Economics and Management Strategy* **10** (2), 173–199.
Frolov D. (2021). Blockchain and institutional complexity: An extended institutional approach. *Journal of Institutional Economics* **17** (1), 21–36.
Giraudo M. (2022). On legal bubbles: Some thoughts on legal shockwaves at the core of the digital economy. *Journal of Institutional Economics* **18** (4), 587–604.
Graf J., Erba J. and Harn R.-W. (2017). The role of civility and anonymity on perceptions of online comments. *Mass Communication and Society* **20** (4), 526–549.
Hardy R.A. and Norgaard J.R. (2015). Reputation in the internet black market: An empirical and theoretical analysis of the deep web. *Journal of Institutional Economics* **12** (3), 515–539.
Harris C. (2018). Institutional solutions to free-riding in peer-to-peer networks: A case study of online pirate communities. *Journal of Institutional Economics* **14** (5), 901–924.
Howell B.E. and Potgieter P.H. (2021). Uncertainty and dispute resolution for blockchain and smart contract institutions. *Journal of Institutional Economics* **17** (4), 545–559.
Johnson R.W., Voorhees C.M. and Khodakarami F. (2023). Is your brand protected? Assessing brand safety risks in digital campaigns. *Journal of Advertising Research* **63** (3), 205–220.
Leitner J.M. (2009). Identifying the problem: Korea's initial experience with mandatory real-name verification on internet portals. *Journal of Korean Law* **9** (1), 83–112.
Lyou M. (2013). The implementation and effects of real name verification policy on Internet in Korea. *Socio-Informatics* **2** (1), 17–29. https://cir.nii.ac.jp/crid/1390282680749737344?lang=en
Mildenberger C.D. (2018). Spontaneous disorder: Conflict-kindling institutions in virtual worlds. *Journal of Institutional Economics* **14** (5), 787–809.
Mildenberger M. and Tingley D. (2019). Beliefs about climate beliefs: The importance of second-order opinions for climate politics. *British Journal of Political Science* **49** (4), 1279–1307.
Norgaard J.R., Walbert H.J. and Hardy R.A. (2018). Shadow markets and hierarchies: Comparing and modeling networks in the Dark Net. *Journal of Institutional Economics* **14** (5), 877–899.
Ohlhaver P. (2025). Community Currencies: The Price of Attention and Cost of Influence in a Networked Age -or-the Price of Entry and Cost of Exit in a Networked Age. SSRN working paper.
Ohlhaver P., Weyl E.G. and Buterin V. (2022). Decentralized Society: Finding Web3's Soul. SSRN Working Paper.
Ostrom E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.
Qian B., Koh T.K. and Zhang X.M. (2024). From anonymity to accountability: How virtual identity disclosure changes the quantity and quality of "likes." Information Systems Research.
Rösner L. and Krämer N.C. (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society* **2** (3), 1–13.
Safner R. (2016). Institutional entrepreneurship, Wikipedia, and the opportunity of the commons. *Journal of Institutional Economics* **12** (4), 743–771.
Santana A.D. (2014). Virtuous or Vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice* **8** (1), 18–33.
Shao C., Ciampaglia G.L., Varol O., Yang K.-C., Flammini A. and Menczer F. (2018). The spread of low-credibility content by social bots. *Nature Communications* **9**, 4787.
Suler J. (2004). The online disinhibition effect. *CyberPsychology and Behavior* **7** (3), 321–326.
Sunstein C.R. and Hastie R. (2015). Garbage in, garbage out? Some micro sources of macro errors. *Journal of Institutional Economics* **11** (3), 561–583.
W3C (2022). Decentralized Identifiers (DIDs) v1.0: Core architecture, data model, and representations. W3C Recommendation 19 July 2022.
W3C (2025). Verifiable Credentials Data Model v2.0. W3C Recommendation 15 May 2025.
Yalcintas A. and Alizadeh N. (2020). Digital protectionism and national planning in the age of the internet: the case of Iran. *Journal of Institutional Economics* **16** (4), 519–536.

Cite this article: Makridis CA (2026). Verified pseudonymity as governance technology: decentralized identity, soulbound tokens, and online discourse. *Journal of Institutional Economics* **22**, e32, 1–23. https://doi.org/10.1017/S1744137426100587

https://doi.org/10.1017/S1744137426100587 Published online by Cambridge University Press