Minimum Structural Guarantees for Identity-Gated ATProto-Compatible Social Systems

High-Level Architecture Proposal (Public Version 20260508-0000-Final)

This model describes minimum guarantees for a freedom-preserving identity-gated overlay. It does not claim that such guarantees can be made universally enforceable through protocol design or cryptography alone. An operator or downstream fork may deliberately bypass them. Accordingly, where cryptography is insufficient, core guarantees must be reinforced through architectural transparency, auditability, documented interoperability rules, institutional separation, and, where necessary, contractual or regulatory constraints.

Identity-gated social systems built on, or alongside, ATProto combine two legitimate but tension-filled objectives. The first is to reduce Sybil abuse, automated manipulation, and mass evasion through verified-human onboarding. The second is to preserve the freedom-preserving properties associated with open social protocols: bounded moderation, meaningful exit, account portability, and the ability to rebuild a discourse presence without being permanently tied to a prior enforcement state. The structural danger does not arise from identity verification as such. It arises when verification is operationalized through a stable backend correlation between verified human identity and public discourse identity, allowing moderation, onboarding, or visibility control to drift from account-level discipline into person-bound exclusion.

1. Problem Statement

An ATProto-compatible system becomes structurally freedom-hostile when a persistent operator-controlled mapping links a verified human to one or more public discourse accounts and remains available to routine operational systems. Once that condition holds, moderation can migrate from account-level intervention to cross-account enforcement, re-registration denial, durable visibility suppression, or long-term exclusion from relevant discourse spaces. The open protocol layer does not itself require this outcome. The risk is introduced by the overlay: identity providers, onboarding gateways, backend correlation stores, moderation systems, AppView indexing decisions, and retention practices. The design objective is therefore not to eliminate moderation or identity assurance, but to ensure that these functions remain structurally bounded, contestable, and resistant to silent person-level escalation.

2. Design Goals

A minimally acceptable identity-gated architecture should satisfy six goals. It should prevent routine person-bound cross-account enforcement. It should prevent irreversible operator-side identity lockout by default. It should ensure that verified identity data is not exposed as a stable moderation-accessible correlation anchor. It should preserve meaningful exit, reset, appeal, and recovery paths. It should impose bounded retention on identity-linked enforcement artifacts. And it should prevent AppView or indexing decisions from becoming opaque, practically irreversible erasure mechanisms. These are not maximalist goals. They do not deny operator safety needs or severe abuse handling. They define the minimum boundary beyond which a verified-human system becomes structurally capable of person-bound exclusion as a routine operating mode.

3. Threat Model Summary

The primary adversaries are the platform operator, the identity provider, the moderation backend, the AppView operator, coordinated user groups, hostile regulators or state actors, and third-party scrapers or archivists. Each can contribute to person-bound exclusion through different paths. The operator can retain backend mappings and deny re-entry. The identity provider can serve as the source of a stable human-bound token. The moderation backend can expand account-level discipline into human-level enforcement. The AppView operator can suppress visibility without formally terminating an account. Coordinated users can weaponize public block and label systems through opaque import practices. External scrapers can preserve public moderation artifacts indefinitely, amplifying harm if operator-side identity correlation exists. The decisive trust boundary lies between verified human identity and public discourse identity. If that boundary fails, downstream protections become fragile even when the underlying protocol remains formally open and federated.

4. Layered Architecture Analysis

At the protocol layer, ATProto provides persistent account identifiers, portable hosting, public records, and delegated moderation structures. These features support integrity, portability, and federation, but they do not themselves require real-world identity correlation. The identity and onboarding overlay is where the structural danger begins. If identity verification produces a stable operator-visible identifier that remains usable after provisioning, the system gains the technical capacity to correlate multiple discourse identities to the same verified human. That capacity may exist even where names, documents, or biometrics are not directly exposed to ordinary services. A persistent internal token is sufficient. The moderation layer must therefore be constrained by design. Ordinary moderation systems should operate only on account-level objects: account identifiers, content identifiers, records, and labels. They should not have routine access to a stable identity correlation layer. The AppView and visibility layer requires separate treatment, because in identity-gated environments visibility is often as consequential as formal access. A user who can technically post but is persistently excluded from discovery, indexing, or mainstream feed inclusion may be functionally exiled. The data-retention layer is equally critical. Correlation artifacts retained under fraud prevention, security, or restricted-processing theories can become the enduring substrate of person-bound enforcement unless they are technically and institutionally fenced off.

5. Illustrative Failure Paths

Three representative failure paths show how exclusion can emerge without explicit protocol changes. First, an onboarding-denial loop arises when a retained correlation token is checked during re-registration and a new account is rejected because a previous account linked to the same human was suspended. Second, an AppView-based exclusion path arises when an account remains technically valid at the PDS layer but is silently de-indexed, heavily downranked, or marked with severe labels by the dominant AppView, making the user practically absent from the relevant discourse sphere. Third, a moderation-escalation path arises when routine moderation systems can query, or indirectly trigger, identity-linked backend state, causing account-level enforcement to become person-level enforcement without an explicit exceptional process. These are abuse paths, not proofs of present implementation; their value lies in showing what the architecture must make difficult by default.

6. Architectural Minimum Guarantees

A viable minimum architecture requires strict separation between identity verification and public discourse identity. Onboarding may confirm human uniqueness, but it must not leave behind a static, backend-visible token that is routinely reusable across account lifecycles. At the high level, this implies context-bound, time-bounded onboarding assertions rather than persistent operator-facing person anchors. In practical design terms, that could be implemented through short-lived, scope-bound credentials, selective-disclosure proofs, or Privacy-Pass-like assertions rather than reusable backend correlation tokens. The system need not prescribe a single cryptographic mechanism, but it must exclude routine reuse of a stable identifier that can trivially map one verified human across multiple accounts. Moderation systems must be restricted to account-level objects by default. Service labels, suspensions, and visibility decisions should apply to the relevant discourse account, not automatically to future accounts associated with the same verified human. Cross-account enforcement, where legally necessary for severe abuse, must be exceptional, manually initiated, narrowly scoped, fully logged, auditable, and subject to higher institutional thresholds than ordinary moderation. Those thresholds should not be rhetorical. They should point to a specific escalatory path, such as judicial mandate, independently reviewed emergency action, or a formally constituted escalation body with an immutable audit trail. Retention and restricted processing must also be technically bounded. It is not sufficient to promise that sensitive correlation data is merely inaccessible in policy terms. Production systems must be designed so that retained identity-linked artifacts are time-bounded, technically segregated, and unavailable to ordinary moderation and onboarding workflows. AppView and labeling power must be transparent and contestable. A dominant AppView in an identity-gated ecosystem can function as the real gatekeeper of discourse even where protocol-level federation remains nominally open. Severe visibility suppression and network-wide exclusion actions must therefore be logged, explainable, reviewable, and appealable. In a federated setting, this cannot be reduced to trust. It requires documented indexing policies, verifiable audit logs, and interoperable disclosure of high-severity visibility actions. One possible future operational path would be a standardized ATProto record, or a functionally equivalent interoperable audit format, for high-severity visibility and de-indexing actions. Finally, recovery and re-entry paths must be meaningful. A

user who leaves voluntarily, deletes an account, or completes a bounded suspension should not be structurally condemned to permanent digital exile merely because the system retains a stable memory of their verified human identity.

7. What Requires Cryptography and What Requires Governance

Some guarantees require cryptographic or protocol-adjacent design. These include minimizing linkability between identity proof and discourse identity, reducing reuse of onboarding assertions, preserving user control over portable account credentials, and ensuring that expired or deleted correlation artifacts become technically unusable. Other guarantees cannot be solved by cryptography alone. AppView fairness, moderation legitimacy, retention discipline, exceptional cross-account enforcement, and resistance to institutional overreach depend on governance, process, organizational separation, auditability, and legal restraint. No cryptographic design can by itself compel a hostile or opportunistic operator, or a downstream fork, to honor these guarantees. Where protocol constraints end, enforceability depends on transparency, auditability, contractual discipline, interoperable review mechanisms, and bounded legal or regulatory controls. A freedom-preserving architecture must therefore assume that some operators and forks will try to reintroduce stable correlation and discretionary exclusion, and it must make those moves visible, contestable, and operationally costly.

8. Trade-Offs, Edge Cases, and Final Principle

No design can simultaneously maximize Sybil resistance, operator safety, abuse prevention, anonymity, and discourse freedom without trade-offs. This proposal does not promise zero risk. It defines a minimum freedom-preserving architecture for systems that insist on identity-gated participation. Severe and repeated abuse remains a hard case. If person-bound cross-account enforcement is ever permitted, it must be treated as an exceptional escalation path rather than a routine moderation convenience. Forks and PDS migrations pose another limit: a user may exit a hostile operator, but that exit collapses if the dominant AppView, or a successor fork, simply recreates the same opaque exclusion logic. Multi-operator federation raises the same challenge in a subtler form. Smaller PDS operators may honor the guarantees while a dominant AppView, relay, or indexing layer quietly nullifies them. The guarantees must therefore operate not only as local operator duties but also as interoperable constraints on visibility governance across operator boundaries. Coordinated user groups remain a residual risk as well. Even where the protocol does not require person-bound persistence, public block artifacts and imported lists can create practical long-tail stigma. This does not justify stronger operator correlation; it reinforces the need to protect the boundary between public discourse history and verified human identity. Future state regulation may introduce further pressure by mandating stronger identity assurance or retention. The final design principle is straightforward. The central danger does not arise from ATProto itself, but from identity-gated operator overlays that retain stable correlation and discretionary control over onboarding, moderation, and visibility. Any acceptable design must make person-bound exclusion structurally difficult, exceptional, auditable, time-bounded, and contestable rather than routine, silent, persistent, and easy to replicate in downstream forks.