DP1 – Federated Authentication & Accountability

The Conditions for Trust — strong, federated authentication and decentralized accountability are the foundations for real trust, safe interaction, and meaningful community.

Purpose of This Draft

This ML-Draft articulates **Desirable Property 1 (DP1)** as a foundational condition for trust in the Meta-Layer. It expands DP1 beyond federated authentication to encompass accountability, adaptive intelligence integration, and foresight-driven governance.

DP1 responds to multiple, overlapping needs:

- The need for decentralized, federated identity without single points of control
- The need for durable accountability without mandatory real-world identity
- The need to govern both human and AI agents coherently
- The need to anticipate predictable abuse and governance failure modes

This draft is intended to guide implementation, governance design, and future ML-RFC development.

Problem Statement

Why identity alone is not enough.

For most of the Web's history, trust has been treated as a byproduct of identity. If a participant could be authenticated, logged in, or verified, trust was assumed to follow. This assumption no longer holds.

At contemporary scale, identity has become cheap to generate, easy to discard, and increasingly decoupled from responsibility. As a result, systems optimized around login and verification routinely fail to protect participants, communities, and institutions from predictable harm.

DP1 begins from a different premise: **trust is not something identity produces on its own**. Trust emerges only when identity is paired with accountability, memory, and governance that operate coherently at the point of interaction.

The Limits of Login-Centric Trust

Login-centric trust models focus on answering a narrow question: who is allowed to enter a system. They do not meaningfully address what happens after entry.

Across platforms and applications, this has produced a recurring pattern:

- Verification is treated as a proxy for good faith
- Identity checks are decoupled from behavior over time
- Enforcement resets when identities are abandoned or recreated
- Predators, scammers, fake accounts, and autonomous agents are able to exploit accumulated trust within an ecosystem before detection or response

Even strong authentication does not prevent abuse when actions are not durably bound to accountable actors. A verified account can still mislead, manipulate, impersonate, or cause harm if there is no persistent relationship between identity and responsibility.

Structural Failure Modes in Today's Web

Several structural conditions compound these weaknesses:

- **Context collapse**: Behavior in one space rarely follows a participant into another
- **No shared memory**: Harm accumulates, but accountability does not
- **Reactive moderation**: Intervention occurs after damage is done
- **Synthetic scale**: Bots, sockpuppets, and automated agents operate faster than human oversight

These are not edge cases. They are systemic properties of the current web. As documented in Meta-Layer research, even the largest platforms remove billions of fake or abusive accounts annually, without meaningfully reducing the underlying incentives or recurrence of abuse.

DP1 as a Shift in Framing

DP1 reframes the problem of trust along three axes:

- From **identity** to **accountability**
- From **platforms** to **zones**
- From **enforcement** to **conditions**

Rather than asking only who someone is, DP1 asks under what conditions participation is allowed, how actions are attributed, and how trust evolves over time.

Threats and Failure Modes

DP1 is not defined in opposition to any single class of actor. Instead, it responds to recurring failure modes that reliably emerge in large-scale, low-friction digital systems.

Two surfaces are in scope. The first is **adversarial**: classes of actor whose behavior reliably degrades trust wherever identity is cheap and accountability is discontinuous. The second is **structural**: the pressures catalogued under *Other Drivers* below, which erode trust even where no

actor intends harm. DP1 treats both as design inputs, because a trust system that defends only against deliberate attackers will still fail under ordinary growth logic.

Scammers

Scammers exploit environments where identity is inexpensive and disposable. Common characteristics include:

- Rapid account creation and abandonment
- Cross-context exploitation of trust signals
- Asymmetric incentives favoring deception

DP1 does not attempt to eliminate scams entirely. Instead, it raises their cost by binding actions to accountable agents, preserving memory across contexts, and enabling communities to escalate trust requirements where appropriate.

Serial Predators and Repeat Abusers

Serial abuse often persists not because it is invisible, but because it is fragmented. When identities reset after enforcement, harm becomes distributed across communities without a durable record.

DP1 addresses this pattern by supporting persistent pseudonymous identity, zone-scoped accountability, and governance mechanisms that allow communities to respond to patterns of harm without resorting to exposure, vigilantism, or centralized surveillance.

Impersonators (Human and AI)

Advances in generative systems have dramatically lowered the cost of impersonation. Voice, image, and text synthesis now allow both humans and AI systems to convincingly misrepresent identity and intent.

DP1 counters impersonation by binding content and actions to verifiable agents, clearly differentiating between human and AI actors, and surfacing provenance signals directly at the interface layer.

Other Drivers (Equally Important)

In addition to explicit abuse, DP1 responds to broader systemic pressures that erode trust even in the absence of malicious intent:

- **Scalable agents operating without visible constraints.** As automation and AI agents scale, they can overwhelm human participation, distort consensus, and accumulate influence faster than human governance processes can respond. Without clear accountability, visibility, and rate limits, both human-operated and autonomous agents can unintentionally or deliberately reshape an ecosystem's trust dynamics. An agent refers to any accountable actor operating in the meta-layer, whether human or AI.

- **Governance capture and conflicts of interest.** Trust systems are vulnerable when those who set or enforce rules have misaligned incentives. Concentrated power, opaque decision-making, or financial dependencies can lead to selective enforcement, uneven accountability, or loss of legitimacy. Over time, this erodes community confidence even if formal rules remain unchanged.
- **Economic models that reward engagement regardless of harm.** Many digital systems optimize for growth, virality, or attention without regard for downstream effects. When visibility, rewards, or influence are tied solely to engagement metrics, deceptive, polarizing, or manipulative behavior is systematically advantaged over constructive participation.

These pressures are structural rather than incidental. They interact and compound one another, producing environments where abuse, manipulation, and trust erosion become predictable outcomes. DP1 is designed to address their combined effects by reshaping incentives, accountability, and governance conditions, rather than treating each pressure in isolation. Taken together, these pressures make clear that trust cannot be repaired solely through backend policy or platform moderation, but must be enacted visibly and continuously at the interface level, where participation, amplification, and accountability actually occur.

Identity-Layer Failure Modes

The mechanisms defined later in this draft each guard against a named failure mode. Collected here, they form the adversarial checklist against which a DP1 implementation should be evaluated:

- **Identity fragmentation**: the same participant appears as unrelated actors across systems, breaking incentives, governance, and trust
- **Identity replay**: actions or credentials are reused across systems to gain unearned trust, rewards, or access
- **Identity laundering**: responsibility is shifted across actors, tools, or delegates to evade accountability
- **Sybil saturation**: large numbers of coordinated identities overwhelm governance, incentives, or visibility systems
- **Semantic drift**: identity signals are assumed to carry the same meaning across contexts where they do not
- **Lineage loss**: identity history cannot be reconstructed, enabling impersonation or erasure of harmful behavior
- **Responsibility laundering**: automated outputs circulate without a traceable, contestable responsible entity
- **Governance drift**: adaptive or automated systems reshape trust conditions without visible human ratification

Each of these is addressed in the sections that follow, and each reappears as a testable condition under *Minimum DP1 Alignment*.

Core Principle

Identity is not merely descriptive at the interface layer; it is the basis for enforceability, continuity, and accountable participation across all Meta-Layer interactions.

Identity is the enforcement boundary of the Meta-Layer. If identity cannot persist across context, delegation, and scale, trust collapses into simulation.

DP1 establishes identity as an enforceable, continuous, and context-bound substrate for all higher-order trust, governance, and interaction within the meta-layer.

Trust in the Meta-Layer emerges when identity, accountability, learning, and foresight are bound together at the interface level.

This principle has several direct implications, each of which is essential to sustaining trust at scale:

- **Identity is plural and contextual, not singular or global.** Participants may operate under different identities in different zones, allowing communities to set appropriate norms without forcing a single global identity model. This preserves inclusion, safety, and local governance autonomy.
- **Accountability attaches to actions, not just names.** Trust depends on the ability to evaluate behavior over time. Binding actions to accountable agents ensures that responsibility persists even when identities are pseudonymous or federated.
- **Memory is preserved without requiring mass surveillance.** Durable, attributable records allow communities to learn from past behavior and prevent repeat abuse, while avoiding continuous monitoring or centralized data collection.
- **Governance adapts over time, but remains human-ratified.** Trust systems must evolve in response to new threats and conditions, yet retain human oversight so that changes remain legitimate, explainable, and aligned with community values.

DP1 does not promise perfect safety or universal trust. Instead, it defines the minimum conditions under which trust can form, persist, and be repaired in complex, multi-actor environments.

Primary Mechanisms and Structural Conditions

DP1 is enacted through a small number of composable mechanism families. None of them is sufficient alone; trust emerges from their combination, and a system that implements some while omitting others will exhibit the failure modes named above.

The mechanism families are:

- **Federated strong authentication** establishes the entry condition. It ensures that access to shared spaces is neither trivially cheap nor monopolized by a single identity authority, while leaving trust decisions to the zone.
- **Sociotechnical zones** translate governance principles into concrete, composable participation rules that operate at the interface, where interaction and amplification actually occur.
- **The identity system layer** carries continuity, integrity, non-transferability, sybil resistance, cross-context semantics, and lineage across environments and time, so that identity remains interpretable as it moves.
- **Action-bound accountability** attaches responsibility to actions rather than names, supported by pseudonymity with responsibility, sealed memory with editability windows, and an explicit trust lifecycle covering escalation, revocation, and recovery.
- **Contestability and due process** make the accountability system itself accountable, through explainability, appeals, and human review at consequential thresholds.
- **Agent classification and asymmetric constraint** keep human, AI, and hybrid participation legible and bounded, and bind automated outputs to responsible entities.
- **Adaptive intelligence under human ratification** allows trust conditions to learn from changing behavior without transferring authority to opaque systems.

These families operate as a stack of conditions rather than a sequence of checks. Authentication without accountability produces verified abuse. Accountability without contestability produces captured enforcement. Zones without continuity produce fragmentation. Adaptation without ratification produces governance drift. The structural condition DP1 imposes is therefore *coherence*: each mechanism must remain legible and enforceable at the point of interaction, and must degrade visibly rather than silently when a boundary is crossed.

The sections that follow define each family in depth.

Federated Strong Authentication (Entry Condition)

Federated strong authentication establishes the baseline condition for participation in the Meta-Layer. Its purpose is not to define trust, but to ensure that entry into shared spaces is not trivially exploitable or monopolized by a single identity authority.

DP1 treats authentication as an **entry condition**, not as a guarantee of trustworthiness or good behavior. Strong authentication reduces frictionless abuse, but only when paired with downstream accountability, memory, and governance does it meaningfully contribute to trust.

Federation as a Baseline Requirement

The Meta-Layer supports federation across multiple identity and authentication systems, including traditional SSO providers, wallets, and emerging credential frameworks. This plural approach ensures:

- No single provider controls access to the Meta-Layer
- Participants can authenticate using systems appropriate to their context
- Communities can adopt stronger or lighter requirements without fragmenting the ecosystem

Federation is essential to resilience. Centralized identity systems concentrate power and risk, while federated systems distribute trust and reduce systemic failure modes.

User-Held Keys and Credentials

Where possible, participants hold their own keys or retain meaningful control over credentials. In cases where custodial systems are used, consent and revocability remain core requirements.

User-held credentials support:

- Participant agency and exit
- Reduced platform lock-in
- Durable accountability across contexts

Authentication Is Not Authorization

Authentication answers the question of *who may enter*. It does not determine *what that participant may do*, *what they may access*, or *how much trust they are afforded*.

All authorization, trust thresholds, and participation rules are defined at the zone level. This separation prevents overloading identity systems with governance logic and keeps trust decisions contextual, transparent, and adaptable.

Sociotechnical Zones as Trust Contexts

Sociotechnical zones are the primary mechanism by which trust conditions are enacted at the interface level. Zones translate abstract governance principles into concrete participation rules that operate where interaction, amplification, and accountability actually occur.

Rather than relying on backend policy enforcement or platform-level moderation alone, zones make trust visible, enforceable, and configurable within the lived experience of participants.

Definition of Sociotechnical Zones

Zones combine technical requirements and social norms to define the conditions under which participation is permitted.

Each zone specifies:

- Acceptable risk levels
- Required proofs or credentials
- Accountability expectations
- Governance and escalation pathways

Communities and applications choose which zones they operate within, allowing trust conditions to vary without fragmenting the underlying Meta-Layer.

Zone-Specific Access Paradigms

Zones represent **orthogonal and composable trust constraints**. Real-world environments typically operate under multiple zones simultaneously, reflecting layered social, legal, and safety requirements.

Open and Identity-Light Zones

- Tokenless open zones
- Pseudonymous zones

Credential and Federation-Based Zones

- Credential-gated zones
- Federated authentication zones

Safety and Constraint-Oriented Zones

- Age-gated zones
- Human-only zones
- High-trust or safety-critical zones

By composing zones, communities can precisely calibrate participation conditions without defaulting to global restrictions.

Compatibility and Exclusion by Design

Zones enforce explicit compatibility requirements. Participation is limited to actors who can meet the defined conditions, making boundaries legible rather than implicit.

This design:

- Prevents silent exclusion or shadow banning
- Makes trust thresholds understandable and contestable
- Allows communities to defend against abuse without universal surveillance

Proof of Humanity as a Zone-Scoped Basis for Participation

Proof of humanity refers to mechanisms that allow a participant to demonstrate that they are a **unique human actor**, without necessarily revealing their real-world identity.

Within DP1, proof of humanity is treated as a **foundational system capability** that must be available across the Meta-Layer, even though its enforcement is zone-scoped and community-defined. This capability is critical at the interface level, where rewards, visibility, and reputation are allocated and where synthetic scale can otherwise distort outcomes.

Some communities may choose to make proof of humanity the basis for participation itself. Others apply it selectively to specific functions such as rewards, governance, rate-limited actions, reputation amplification, or access to safety-critical spaces.

Key principles include:

- Proof of humanity is available system-wide, but enforced at the zone level
- Communities decide when and how proof of humanity is required, including making it a prerequisite for participation, rewards, governance, or amplification
- Proof of unique humanity is commonly required to unlock rewards, boost virality, or affect transferable reputation, where synthetic scale would otherwise overwhelm human participation
- Multiple proof mechanisms may coexist, be combined, or be phased out over time
- Proof of humanity does not imply real-name identity or permanent disclosure
- Requirements may be stricter in high-trust, safety-critical, or resource-allocation contexts

By treating proof of humanity as an enduring and adaptable capability rather than a fixed mechanism, DP1 enables long-term defense against synthetic scale and impersonation while preserving pluralism, pseudonymity, and local governance autonomy.

Identity System Layer: Continuity, Integrity, and Adversarial Resilience

Beyond authentication and zone-scoped participation, DP1 requires a coherent identity system layer that persists across environments, interactions, and time. This layer is the enforcement boundary of the Meta-Layer. If identity cannot maintain continuity under scale, delegation, and interoperability, trust collapses into simulation.

The identity system layer ensures that actions, reputation, and accountability remain meaningfully bound to agents even as they move across zones, tools, and contexts.

Identity Continuity Across Systems

Identity must persist across platforms, zones, and applications without fragmenting into unrelated entities.

Continuity requires:

- Stable identifiers or linkable identity references
- Preservation of accountability history across environments
- Explicit signaling when identity properties degrade or reset

A failure mode is **identity fragmentation**, where the same participant appears as unrelated actors across systems, breaking incentives, governance, and trust.

Identity Integrity and Anti-Replay Guarantees

Identity-bound actions must not be duplicable across systems without attribution.

This requires:

- Binding actions to identity with verifiable provenance
- Preventing replay of actions or credentials across contexts
- Ensuring that contributions cannot be re-used to extract value multiple times

A failure mode is **identity replay**, where actions or credentials are reused across systems to gain unearned trust, rewards, or access.

Identity Non-Transferability and Delegation Boundaries

Identity must not be freely transferable in ways that detach responsibility from the original actor.

Delegation is permitted, but must be:

- Explicit
- Scoped
- Attributable

This ensures that actions taken by agents, tools, or collaborators remain traceable to accountable principals.

A failure mode is **identity laundering**, where responsibility is shifted across actors to evade accountability.

Resistance to Sybil and Coordinated Identity Attacks

Identity systems must make large-scale duplication, coordination, or synthetic amplification detectable, constrained, or economically costly.

Mechanisms may include:

- Proof-of-humanity where appropriate
- Behavioral analysis and rate limits
- Reputation weighting and trust decay

A failure mode is **sybil saturation**, where large numbers of coordinated identities overwhelm governance, incentives, or visibility systems.

Cross-Zone Identity Semantics and Degradation

Identity does not carry identical meaning across all zones.

Systems must:

- Signal when identity guarantees change across contexts
- Prevent misinterpretation of reputation or credentials
- Allow communities to define how identity signals are interpreted locally

A failure mode is **semantic drift**, where identity signals are incorrectly assumed to carry the same meaning across different contexts.

Identity Memory and Lineage

Identity must retain a reconstructable history of actions, credentials, and governance interactions over time.

Lineage enables:

- Attribution of contributions
- Detection of behavioral patterns
- Accountability across time and context

Breaks in lineage must be treated as risk signals rather than neutral events.

A failure mode is **lineage loss**, where identity history cannot be reconstructed, enabling impersonation or evasion.

This identity system layer does not require centralization or global identity unification. It requires coherence. Identity must remain usable, accountable, and interpretable across the Meta-Layer without collapsing into surveillance or fragmentation.

Accountability as a First-Class Property

Accountability is the core mechanism through which trust becomes durable in the Meta-Layer. While authentication governs entry, accountability governs behavior over time. Without it, trust signals decay, abuse repeats, and governance loses legitimacy.

DP1 treats accountability as a first-class property that operates continuously at the interface level, binding actors to their actions in ways that are visible, attributable, and contestable, without requiring real-world identity disclosure.

Action-Bound Accountability

In the Meta-Layer, accountability attaches to actions, not merely to identities. Every meaningful action, such as posting content, issuing judgments, triggering automation, or influencing visibility, is bound to an accountable agent identifier.

This ensures that:

- Actions have clear provenance
- Responsibility persists across time and context
- Trust assessments can be based on behavior, not credentials alone

Action-bound accountability allows communities to reason about patterns of conduct without collapsing participation into real-name systems or centralized surveillance.

Pseudonymity with Responsibility

DP1 explicitly supports pseudonymous participation, recognizing its importance for safety, expression, and inclusion. Pseudonymity, however, does not imply anonymity from accountability.

Persistent pseudonymous identities allow participants to:

- Build reputation over time
- Be held responsible for repeated behavior
- Participate across zones without exposing real-world identity

Communities may permit multiple personas per participant, subject to local rules, provided that accountability requirements are met. This balances flexibility with responsibility, enabling participation without enabling evasion.

Sealed Memory and Editability Windows

To balance forgiveness, accuracy, and integrity, DP1 supports time-bound editability followed by sealing.

Participants may edit or retract contributions within community-defined windows. After this period, contributions become sealed: immutable, attributable, and part of the shared civic memory.

Sealed memory:

- Prevents retroactive manipulation
- Preserves historical context
- Enables learning from past behavior

Communities may determine whether edit histories are retained, visible, or restricted, but the existence of durable memory is essential for trust to accumulate.

Trust Lifecycle, Revocation, and Recovery

Trust in the Meta-Layer is not binary. It evolves.

Zones define explicit conditions for:

- **Escalation and intervention**: graduated responses to harmful or destabilizing behavior
- **Temporary suspension or restriction**
- **Revocation of access or privileges**
- **Reinstatement or recovery**

Revocation is zone-scoped by default, avoiding unnecessary global punishment. Memory persists across decisions.

Contestability, Appeals, and Due Process

For accountability systems to be trusted, they must themselves be accountable. DP1 therefore treats contestability and due process as essential trust infrastructure, not optional governance overhead.

Participants must be able to understand, challenge, and appeal decisions that materially affect their participation, visibility, reputation, or access.

Key principles include:

- **Transparency**: Decisions affecting trust or access must be explainable and grounded in visible rules or signals.
- **Contestability**: Participants must have mechanisms to dispute actions taken against them.
- **Human Oversight**: Escalation thresholds require human review, particularly where consequences are significant.
- **Explainable Automation**: AI-assisted flagging or enforcement must be intelligible to affected parties.

Appeals processes reinforce legitimacy. They help communities detect governance failure, correct errors, and adapt rules over time.

By embedding contestability directly into trust systems, DP1 ensures that accountability strengthens trust rather than undermining it.

In summary:

- Accountability systems must be challengeable and auditable
- Participants must be able to contest decisions affecting access, reputation, or visibility
- Escalation thresholds require human review
- AI-assisted flagging and enforcement must be explainable

- Appeals are treated as trust infrastructure, not optional overhead

---

Human and AI Agents Under DP1

DP1 treats both human and artificial agents as first-class participants in the Meta-Layer, while recognizing that they differ fundamentally in capacity, scale, intent, and risk profile. Trust cannot be sustained if these differences are ignored, obscured, or flattened.

The goal of DP1 is not to exclude AI agents categorically, but to ensure that their participation is **legible, bounded, and accountable** in ways that preserve human agency and community governance.

Agent Classification and Visibility

An *agent* refers to any actor capable of taking actions that affect shared environments, visibility, reputation, or outcomes within the Meta-Layer.

DP1 requires clear classification between:

- Human agents
- AI or automated agents
- Hybrid or assisted agents, where human intent is mediated by automation

This classification must be **visible at the interface level**, allowing participants to understand whether they are interacting with a human, an AI system, or a combination of both. Hidden or ambiguous agent identity erodes trust and enables manipulation.

Symmetric Accountability, Asymmetric Constraints

DP1 applies accountability symmetrically: all agents are accountable for their actions. However, constraints are applied asymmetrically, reflecting differences in scale, speed, and potential impact.

For example:

- AI agents may be subject to stricter rate limits, scope restrictions, or amplification caps
- Certain zones may restrict participation to human agents only
- Higher proof thresholds may apply where AI activity could distort consensus, rewards, or governance

This approach avoids both extremes: granting AI agents unchecked parity with humans, or exempting them from accountability altogether.

Binding AI Outputs to Responsible Entities

AI agents do not operate independently of human or institutional responsibility. DP1 therefore requires that AI outputs be bound to a responsible entity, such as:

- The operator deploying the agent
- The organization maintaining it
- A community governance structure authorizing its use

In high-trust or safety-critical zones, anonymous autonomous agents are not permitted. Responsibility must be traceable, contestable, and enforceable.

By binding AI behavior to accountable entities, DP1 prevents responsibility laundering while enabling beneficial automation under governed conditions.

Adaptive Intelligence Integration (RLADP)

Static trust and governance systems degrade over time. Incentives shift, adversaries adapt, and behaviors drift. DP1 therefore anticipates the need for adaptive intelligence to support, but not replace, human and community governance.

Why Static Governance Fails

At scale, purely static rules and manual moderation encounter predictable limits:

- Human moderators cannot match the speed or volume of adversarial behavior
- Rules ossify and become misaligned with lived practice
- Bad actors learn to game fixed thresholds and heuristics

Without adaptation, governance systems either become overly permissive or increasingly brittle.

RLADP as Advisory Infrastructure

DP1 envisions adaptive intelligence, including reinforcement learning and approximate dynamic programming (RLADP), as **advisory infrastructure**.

Adaptive systems may:

- Detect emerging patterns of abuse or manipulation
- Surface signals about shifting norms or risk profiles
- Propose adjustments to thresholds, friction, or zone parameters

They may not:

- Unilaterally change rules
- Impose sanctions without human ratification
- Operate as opaque or unchallengeable authorities

Transparency and Auditability

All adaptive processes must be observable and auditable. Communities must be able to understand:

- What signals are being used
- How recommendations are generated
- What effects adaptations have produced

This visibility is essential to preventing hidden governance drift and maintaining legitimacy.

Human and Community Ratification

Adaptive intelligence proposes; humans decide.

Material changes to trust conditions, enforcement thresholds, or governance rules require explicit human or community ratification, using processes appropriate to the zone.

By constraining adaptive intelligence within transparent, ratified loops, DP1 enables learning without surrendering agency or accountability.

Governance, Accountability, and Agency Surfaces

DP1 is not satisfied by identity architecture alone. The conditions for trust must be *experienced* as navigable surfaces, or they remain private properties of a backend that participants are asked to take on faith. A system can bind every action to an accountable agent and still fail DP1 if participants cannot see the binding, cannot tell which zone rules apply to them, and cannot challenge a decision that changes their standing.

Participants must be able to:

- see which zone they are currently participating in, what proofs it requires, and which of those proofs they currently satisfy
- see whether the actor they are interacting with is classified as human, AI, or hybrid, and which entity is responsible for that actor's behavior
- inspect what is attributed to their own agent identifier, and within what editability window a contribution remains revisable before it is sealed
- understand their current trust state in a zone, including any active restriction, its stated basis, its duration, and the recovery pathway
- initiate an appeal against a decision affecting access, visibility, or reputation, and observe its status rather than waiting on an invisible queue
- understand when identity continuity is about to degrade — on export, migration, fork, or persona separation — before the degradation occurs

Communities must be able to:

- define, publish, and revise the zone conditions under which they operate, including proof requirements and escalation thresholds
- see who currently holds enforcement authority in the zone, under what mandate, and for how long
- audit whether enforcement actions were executed as ratified, and whether adaptive recommendations were adopted, modified, or rejected
- review aggregate accountability outcomes — appeal volumes, reversal rates, escalation frequency — without exposing individual participants
- preserve governance memory of why a trust rule exists, what incident prompted it, and what effect it produced
- fork zone rules, or exit, when legitimacy breaks down, without stripping members of identity continuity

Example: A participant restricted from amplifying content in a high-trust zone sees the specific rule invoked, the signal that triggered it, whether a human ratified the action, the date the restriction lapses, and a single path to contest it. The community, separately, can see how often that rule fires and how often contests succeed.

Without these surfaces, accountability becomes indistinguishable from arbitrary enforcement. The failure mode is **opaque legitimacy**, where a system's trust decisions may in fact be sound but cannot be verified as such, and therefore accumulate distrust at the same rate as capture would.

Incentives and Power Analysis

Identity systems are never neutral infrastructure. Whoever controls issuance, verification, revocation, or resolution of identity holds structural power over participation itself, and that power is valuable enough that it will be sought whether or not the formal rules permit it.

DP1 therefore treats incentive structure as part of the trust surface, not as context outside it.

Several pressures recur:

- **Verification as a rent.** Where a single provider becomes the de facto path to credible participation, verification becomes a tollgate. Federation is not merely a resilience property; it is an anti-rent property, ensuring that no issuer can price access to the Meta-Layer.
- **Engagement-weighted visibility.** Where reach is allocated by engagement alone, synthetic scale is economically rational. Proof-of-humanity gating for rewards and amplification is therefore an incentive intervention, not only a security control.

- **Enforcement asymmetry.** Those who set rules, enforce them, and benefit from their outcomes are frequently the same actors. DP1 requires that these relationships be visible (see *Conflict of Interest (COI) Visibility*), because conflicts of interest degrade legitimacy even absent bad intent.
- **Reset economics.** When identity is cheap to abandon, sanctions become a cost of doing business rather than a deterrent. Persistent identity, lineage, and privilege loss on reset change the arithmetic of abuse.
- **Data leverage from accountability.** Accountability infrastructure generates exactly the records that surveillance economies value. DP1's insistence on pseudonymity, zone-scoping, and non-propagation by default is an incentive constraint on the operators of trust systems themselves.
- **Automation subsidy.** Automated agents can absorb friction that would deter humans. Asymmetric rate, scope, and amplification constraints prevent capital from purchasing disproportionate presence.

Example: A platform requires proof of humanity for rewards but exempts partner integrations to protect a revenue relationship. No rule changed, yet the incentive to route abuse through partner surfaces was created.

Why this matters: A trust system that does not act on incentives will be routed around by them. DP1 therefore expects that materially relevant incentives — who funds enforcement, who profits from visibility, who benefits from a given verification path — be legible to communities, and that zones retain the ability to constrain those incentives locally.

Community Signals Informing DP1

DP1 reflects recurring themes from community submissions, workshops, and discussions across the Meta-Layer initiative.

While individual inputs vary, several consistent signals emerge:

- **A strong preference for pseudonymity with accountability, rather than forced real-name identity.** Communities repeatedly emphasize the need to separate accountability from exposure. Participants want the ability to speak, contribute, and organize without tying activity to real-world identity, while still ensuring that actions carry responsibility over time. This reflects lived experience in environments where real-name policies create safety risks, suppress participation, or concentrate power, without reliably preventing abuse.
- **Frustration with repeat abusers enabled by identity resets and fragmented enforcement.** Many communities report that harmful behavior persists not because it goes unnoticed, but because enforcement lacks continuity. When identities can be cheaply abandoned and re-created, sanctions lose meaning and abuse becomes a cost of

doing business. This signal directly informs DP1's emphasis on persistent, pseudonymous identity and durable memory.

- **Concern about synthetic scale, including bots and AI agents overwhelming human participation.** Participants consistently describe environments where automated or semi-automated agents dominate visibility, rewards, or discourse. Even benign automation can distort outcomes when scale is unchecked. This concern motivates proof-of-humanity capabilities, asymmetric constraints for AI agents, and explicit limits on amplification and rate.
- **Distrust of opaque or centralized moderation and fear of governance capture.** Communities express low confidence in trust systems where rules are enforced invisibly or controlled by unaccountable actors. Perceived capture, selective enforcement, or undisclosed incentives erode legitimacy even when formal policies appear sound. This drives DP1's requirements for transparency, contestability, and visible governance at the interface level.
- **Desire for preventative and restorative approaches rather than purely punitive systems.** Many submissions emphasize that punishment alone does not build trust. Communities want mechanisms that prevent harm upstream, allow for learning and repair, and support reintegration where appropriate. This signal underlies DP1's focus on graduated responses, recovery pathways, and governance that evolves through foresight rather than crisis.

These signals reinforce the core framing of DP1: trust must be designed as a set of conditions that balance agency, safety, and legitimacy, rather than imposed through static rules or centralized control. DP1 is therefore best understood not as a single solution, but as a shared response to patterns of failure repeatedly identified by communities operating at scale.

Foresight and Failure Design

DP1 treats foresight not as speculation, but as a core governance discipline. Large-scale sociotechnical systems fail in recognizable ways. When trust systems are designed only for normal operation, they become brittle under stress, capture, or adversarial pressure.

Minefield thinking refers to the practice of deliberately anticipating where incentives, power, and scale are likely to produce failure, and designing safeguards in advance rather than reacting after harm has occurred.

Governance as Anticipatory Design

Most trust failures are not surprises. They arise from known dynamics such as incentive misalignment, asymmetric power, scale effects, and adversarial learning.

DP1 therefore treats governance as an anticipatory design problem. Communities are encouraged to:

- **Identify foreseeable abuse and failure modes.** Rather than assuming good-faith participation as a default, communities are encouraged to explicitly map how their systems could be exploited, stressed, or captured at scale. This includes considering adversarial behavior, incentive misalignment, power concentration, and unintended consequences of well-meaning rules.
- **Encode preventative friction rather than relying solely on punishment.** Preventative friction includes rate limits, proof thresholds, graduated permissions, and contextual checks that slow or deter harmful behavior before it escalates. This reduces reliance on after-the-fact enforcement, which is often costly, contentious, and insufficient to prevent harm.
- **Periodically reassess assumptions as conditions change.** Trust systems operate in dynamic environments. Communities are encouraged to revisit governance assumptions as participation grows, technologies evolve, or incentives shift, ensuring that rules remain aligned with lived practice rather than ossifying over time.

This approach shifts governance from reactive moderation to continuous risk management.

Conflict of Interest (COI) Visibility

Trust erodes when participants cannot see whose interests shape rules and enforcement. DP1 requires that material conflicts of interest be surfaced structurally rather than assumed away.

This includes visibility into:

- **Funding sources and economic incentives.** Communities benefit from understanding who funds infrastructure, moderation, or tooling, and how revenue models or token incentives may shape decision-making. Visibility into economic incentives helps participants evaluate whether rules are aligned with collective goals or subtly optimized for growth, extraction, or control.
- **Governance authority and decision rights.** Trust depends on knowing who has the power to set rules, enforce them, and change them over time. Clear articulation of decision rights allows participants to assess legitimacy, understand escalation pathways, and distinguish community governance from operator discretion.
- **Relationships between operators, enforcers, and beneficiaries.** When the same actors design rules, enforce them, and benefit from their outcomes, conflicts of interest can arise even without malicious intent. Making these relationships explicit allows communities to surface bias, challenge capture, and adjust governance structures before trust erodes.

By making incentives legible, communities can better assess legitimacy, detect capture early, and sustain confidence in governance over time.

Governance Pre-Mortems

DP1 encourages communities to conduct periodic governance pre-mortems: structured exercises that ask how current rules or systems might fail under plausible future conditions.

Pre-mortems may examine:

- **How rules could be gamed at scale.** Communities are encouraged to consider how well-intentioned rules might be exploited when participation grows, automation increases, or incentives shift. This includes identifying loopholes, edge cases, or feedback loops that could advantage bad-faith actors.
- **Where enforcement might become selective or biased.** Pre-mortems surface the risk that enforcement could drift toward favoritism, uneven application, or disproportionate impact on certain groups. Making these risks explicit allows communities to design checks, audits, or appeals in advance.
- **How new technologies or actors could distort participation.** Emerging tools, AI capabilities, or new classes of participants may change how power and influence are exercised. Pre-mortems help communities anticipate these shifts rather than reacting after harm occurs.

The goal is not prediction, but preparedness. Pre-mortems create shared awareness of fragility, normalize course correction, and reduce the social and political cost of adaptation.

Exit, Fork, and Kill Switches

No governance system should assume its own permanence. DP1 treats exit as a safety feature rather than a failure, recognizing that the ability to leave or disengage is essential to legitimacy.

Communities and participants should have:

- **Clear paths to exit without losing identity or accountability continuity.** Participants should be able to leave a space without being erased or forced to abandon their history, allowing accountability and learning to persist across contexts.
- **The ability to fork governance or norms when consensus breaks down.** When irreconcilable differences emerge, forking allows communities to diverge without coercion, preserving agency while limiting destructive conflict.
- **Emergency mechanisms to pause or disable systems causing systemic harm.** Kill switches or pauses provide a last-resort safeguard against cascading failure, runaway automation, or captured governance.

These safeguards limit the blast radius of governance failure, reduce incentives for capture, and make participation safer by design.

Cross-Zone Failure Containment

In a multi-zone environment, failures should be contained by default. DP1 assumes that trust loss, enforcement actions, and reputational signals are local unless explicitly propagated.

Communities define:

- **When signals remain zone-scoped.** Localizing consequences prevents minor or context-specific failures from unfairly affecting participation elsewhere.
- **When and how signals may propagate across zones.** Communities may choose to share certain signals across zones where risks overlap, but such propagation should be deliberate, transparent, and governed.
- **What thresholds justify broader impact.** Explicit thresholds help distinguish between isolated incidents and systemic harm, enabling proportional response.

This containment prevents cascading harm while preserving the ability to respond proportionally to serious or systemic abuse.

Failure is expected. Silent failure is not. A DP1-aligned system is one in which the degradation of trust conditions — a broken lineage, an unratified adaptation, an unresolved appeal backlog, a zone quietly loosening its proof requirements — is observable while it is still correctable.

Relationship to Other Desirable Properties

DP1 is foundational and cross-cutting. Many other Desirable Properties depend directly on the conditions it establishes.

In particular:

- **Properties related to safety, harm reduction, and abuse prevention** rely on durable accountability, shared memory, and zone-scoped enforcement to function effectively. Without persistent attribution and the ability to recognize patterns of behavior over time, safety mechanisms degrade into reactive moderation that fails to prevent repeat harm or coordinated abuse.
- **Properties concerning agency, consent, and autonomy** depend on federated identity, user-held credentials, and contestable governance so that participants can meaningfully choose how they engage, under what conditions, and with which authorities. Without these foundations, consent becomes nominal, exit becomes costly, and power asymmetries harden.
- **Properties addressing AI participation, automation, and alignment** require clear agent differentiation, asymmetric constraints, and binding accountability to prevent synthetic scale from overwhelming human judgment or distorting collective outcomes. DP1 establishes the conditions under which AI systems can participate without eroding trust or legitimacy.
- **Properties focused on collective intelligence, coordination, or governance** assume the existence of trustworthy participation, legible authority, and adaptive learning loops. Collective sensemaking and coordination cannot emerge where participants doubt who

is acting, how decisions are made, or whether systems can learn from failure without capture.

More specifically:

- **DP2** supplies the levers that accountable actors hold. DP1 without DP2 produces accountability without agency, which is surveillance; DP2 without DP1 produces agency without attribution, which is impunity.
- **DP3** governs how trust conditions themselves change. Zone rules, escalation thresholds, and enforcement mandates are governance artifacts subject to DP3's tiering, revocability, and memory requirements.
- **DP4** bounds what the accountability layer may retain, infer, or transfer. Identity lineage must be reconstructable without becoming a surveillance corpus.
- **DP5** provides the namespace substrate. Persistent identifiers, personal identifiers, and controller bindings are what make identity continuity addressable rather than notional.
- **DP6** and **DP9** depend on anti-replay and sybil resistance, because value transfer and contribution rewards are the surfaces on which duplicated identity pays best.
- **DP7** determines whether identity guarantees survive movement between systems, or degrade silently at integration boundaries.
- **DP11**, **DP12**, and **DP13** extend agent classification and asymmetric constraint into disclosure, policy execution, and containment for automated actors.
- **DP14** and **DP15** carry provenance and transparency for the records on which accountability rests.
- **DP20** determines who ultimately owns and can fork the identity and enforcement infrastructure, which is the final backstop against capture.

Weakness or ambiguity in DP1 propagates upward, undermining the effectiveness of other properties. Conversely, a strong DP1 enables the Meta-Layer to support more advanced coordination, safety, and governance capabilities without reverting to centralized control.

Non-Goals and Explicit Boundaries

DP1 deliberately defines the *conditions* for trust rather than attempting to solve all problems associated with identity, abuse, or governance on the internet. Explicitly stating non-goals is essential to prevent scope creep, misinterpretation, and inappropriate application of this property.

DP1 does **not** attempt to:

- **Enforce real-name policies globally.** DP1 explicitly rejects the assumption that real-world identity disclosure is a prerequisite for trust. Mandatory real-name systems often

increase risk, suppress participation, and centralize power without reliably preventing abuse.

- **Eliminate all abuse or deception.** No trust system can guarantee perfect safety. DP1 focuses on raising the cost of harm, preserving accountability, and enabling learning and repair, rather than promising total prevention.
- **Centralize identity, moderation, or governance.** DP1 is incompatible with architectures that concentrate authority in a single platform, provider, or enforcement body. Trust is treated as a plural, zone-scoped property rather than a global control mechanism.
- **Replace legal systems or law enforcement.** DP1 operates at the interface and governance layer. It does not supersede legal processes, nor does it attempt to adjudicate crimes or enforce jurisdictional law.

DP1 also does not:

- **Mandate a single proof-of-humanity mechanism.** The capability must be available system-wide; the mechanism remains a community choice, and multiple mechanisms may coexist or be superseded.
- **Establish a global reputation score.** Trust signals are zone-scoped by default. Cross-zone propagation is an explicit, governed act, not an emergent property of participation.
- **Prohibit anonymity.** Unattributed action may exist. What DP1 constrains is its ability to affect shared resources, rewards, governance, or amplification.
- **Treat delegation as a transfer of responsibility.** Agents, tools, and collaborators may act on a participant's behalf, but the accountable principal does not change by virtue of delegation.
- **Guarantee that identity guarantees survive every boundary.** Where continuity, provenance, or trust semantics degrade, DP1 requires that the degradation be signaled, not that it never occur.

Failure mode: **overreach**, where DP1 is invoked to justify mandatory identification, global enforcement, or surveillance infrastructure that its own non-goals exclude.

By naming these boundaries explicitly, DP1 remains adaptable across cultures, legal regimes, and communities, while resisting overreach or misuse.

Minimum DP1 Alignment (Non-Normative)

Minimum alignment is not a feature checklist. It is the threshold at which an identity system can be considered **enforceable, portable, and resistant to trivial abuse**.

A system that does not meet these conditions may function, but it cannot reliably sustain trust under scale, automation, or adversarial pressure.

At minimum, a system claiming alignment with DP1 must satisfy the following **irreducible conditions**:

Persistent Identity Continuity

- Actions MUST be bound to a persistent agent identifier that survives across sessions and contexts
- Identity resets MUST be rate-limited, detectable, or carry loss of accumulated privileges
- Systems MUST signal when identity continuity is broken or degraded

Failure mode: identity reset cycles that enable repeated exploitation of incentives and governance.

Verifiable Action Attribution

- All meaningful actions (content, transactions, moderation, automation) MUST be attributable to an agent
- Attribution MUST include sufficient provenance to audit origin and context
- Anonymous or unattributed actions MAY exist, but MUST be constrained from affecting shared resources or governance

Failure mode: untraceable actions that erode accountability and enable manipulation.

Anti-Replay and Non-Duplication Guarantees

- Identity-bound actions and credentials MUST NOT be reusable across systems without explicit lineage
- Systems MUST detect or prevent replay of contributions, credentials, or proofs
- Cross-system transfers MUST preserve attribution or clearly signal loss of guarantees

Failure mode: replay attacks that extract duplicate rewards, access, or influence.

Sybil Resistance Under Scale

- Systems MUST impose friction or constraints that make large-scale identity duplication costly or ineffective
- Mechanisms MAY include proof-of-humanity, rate limits, staking, or reputation weighting
- Systems MUST remain functional under coordinated identity attacks

Failure mode: sybil saturation overwhelming incentives, governance, or visibility.

Zone-Scoped Enforcement and Revocation

- Enforcement actions MUST be defined and executed within explicit zones
- Revocation, restriction, and recovery pathways MUST be visible and rule-based

- Global or opaque enforcement MUST NOT be the default

Failure mode: arbitrary or centralized enforcement that undermines legitimacy.

Human and AI Agent Differentiation

- Systems MUST visibly distinguish between human, AI, and hybrid agents at the interface level
- Automated agents MUST be subject to appropriate constraints (rate, scope, amplification)
- AI outputs MUST be bound to a responsible entity

Failure mode: indistinguishable agents enabling manipulation, impersonation, and synthetic dominance.

Identity Lineage and Memory

- Systems MUST maintain reconstructable identity lineage for actions, credentials, and governance events
- Breaks in lineage MUST be treated as risk signals, not neutral transitions
- Historical records MUST be tamper-resistant after defined edit windows

Failure mode: lineage loss enabling impersonation, laundering, or erasure of harmful behavior.

Contestability and Due Process Baseline

- Participants MUST have access to mechanisms for contesting enforcement actions
- Appeals pathways MUST exist for decisions affecting access, visibility, or reputation
- High-impact decisions MUST include human review or ratification

Failure mode: unchallengeable systems that degrade into opaque or captured governance.

These conditions define the **minimum viable enforcement layer for identity** in the Meta-Layer.

Partial implementations that omit continuity, attribution, anti-replay guarantees, or sybil resistance SHOULD NOT be considered aligned with DP1, regardless of authentication strength or interface design.

Open Questions and Future Work

DP1 establishes foundational conditions for trust, but it does not resolve all questions required for long-term interoperability, standardization, and global deployment. The following areas are intentionally left open for further research, experimentation, and community deliberation.

- **Standardization of proof-of-humanity mechanisms.** While DP1 requires the availability of proof of unique humanity as a system capability, it does not prescribe specific mechanisms. Multiple decentralized approaches already exist, such as Fractal ID and other proof-of-humanity systems, each with different tradeoffs around privacy, accessibility, cost, and resistance to gaming. Communities may choose the mechanisms that best fit their norms and risk profiles. Over time, however, it is likely that one or a small number of widely trusted proof-of-humanity systems will emerge for use at the Meta-Layer or Overweb level, providing a common baseline that communities are encouraged, but not required, to adopt. Open questions include how such proofs can remain privacy-preserving, globally accessible, and adaptable over time, as well as how multiple proof systems might interoperate, be bridged, or be composed.
- **Cross-zone reputation portability.** DP1 assumes that accountability and trust signals are zone-scoped by default. Further work is needed to determine when and how reputation, sanctions, or trust signals should move across zones without creating unjust spillover effects or de facto global scoring systems.
- **Liability and responsibility models for autonomous agents.** As AI agents become more capable and autonomous, clearer models are needed for assigning responsibility across operators, deployers, tool providers, and communities. DP1 establishes binding to responsible entities, but does not yet resolve how liability should be apportioned in complex, multi-actor systems.
- **Thresholds for escalation across zones.** Communities will need shared patterns for deciding when local failures warrant broader response. This includes defining proportional thresholds, evidentiary standards, and governance processes for cross-zone escalation without undermining local autonomy.
- **Meta-Layer coordination mechanisms.** DP1 implies the need for shared coordination mechanisms capable of maintaining coherent context across human participants, AI agents, content objects, and communities. Such mechanisms would support converging profiles, scoped capabilities, and durable accountability without requiring centralized control. Emerging approaches to structured context exchange and agent coordination, such as Model Context Protocols (MCP), may offer useful design patterns for this layer when generalized beyond model runtime to sociotechnical actors. The design of any such Meta-Layer coordination protocol remains an open area of research and standardization.

Further questions concern how a trust system is measured and how its guarantees behave under pressure:

- how to measure trust-system health beyond enforcement counts, for example appeal reversal rates, time-to-resolution, and the frequency of unratified adaptive changes
- how to represent identity continuity degradation in interfaces that ordinary participants can interpret at the moment it matters
- how to support legitimate persona separation without enabling evasion, given that both look similar from the outside
- how to preserve accountability lineage while honoring data minimization and deletion expectations under DP4
- how to sustain proof-of-humanity accessibility for participants without documents, stable devices, or reliable connectivity

These questions are not gaps in DP1, but signals of where future ML-Drafts and ML-RFCs may be required as the Meta-Layer matures.

Path Toward ML-RFC

This ML-Draft is intended as exploratory scaffolding rather than a finalized specification. Progression toward an ML-RFC should be guided by rough consensus, iterative refinement, and practical validation.

Key steps toward ML-RFC status include:

- **Soliciting broad community review and critique.** Feedback from implementers, civil society, governance practitioners, and researchers is essential to test assumptions and surface edge cases.
- **Identifying points of rough consensus.** Not all aspects of DP1 must be settled to advance. Emphasis should be placed on stabilizing core invariants such as action-bound accountability, zone-scoped trust, and contestability.
- **Clarifying implementation invariants.** Future drafts should distinguish clearly between invariant requirements and flexible design space, reducing ambiguity for builders while preserving pluralism.
- **Separating exploratory elements from normative commitments.** Concepts such as coordination protocols or adaptive intelligence should mature through dedicated drafts before being incorporated normatively.
- **Promoting stable elements to ML-RFC status.** Once sufficient consensus and operational understanding exist, portions of DP1 may be advanced as ML-RFCs to serve as durable reference points for the Meta-Layer ecosystem.

Graduation criteria SHOULD include demonstrable evidence, not description alone:

- conformance tests for each condition under *Minimum DP1 Alignment*, including adversarial scenarios for replay, sybil saturation, and lineage loss
- reference schemas for agent identifiers, attribution records, zone profiles, and enforcement receipts
- evidence that identity continuity survives migration into at least one independent system, or that degradation is signaled where it does not
- evidence that appeals resolve within stated timelines and that reversal outcomes are observable
- evidence that adaptive recommendations were ratified, modified, or rejected by humans with a visible record

This progression reflects the Meta-Layer's commitment to transparency, accountability, and participatory standards development.

Closing Orientation

DP1 is the claim that trust can be built without requiring people to surrender who they are.

It refuses the two familiar bargains of the current web. The first offers safety in exchange for identification, and delivers surveillance. The second offers freedom in exchange for accountability, and delivers impunity. DP1 asserts that neither trade is necessary: identity can be plural and pseudonymous while actions remain durably attributable, and communities can enforce their own conditions without a central authority defining who counts.

When DP1 is strong, participation carries weight. Reputation accumulates because it cannot be trivially reset. Automation is welcome because it is legible and bounded. Enforcement is credible because it can be contested. Communities can set their own thresholds because zones make difference composable rather than fragmenting.

When DP1 is weak, everything above it becomes performance. Governance votes that anyone can flood. Agency controls over flows no one can attribute. Ethical AI commitments with no responsible entity behind an output. Commerce where the same identity collects twice.

DP1 defines the conditions under which trust can emerge. Without it, the meta-layer becomes another surface. With it, the meta-layer becomes a place.
