DP11 – Safe and Ethical AI

The Meta-Layer makes AI transparent, explainable, and aligned with human values and community goals.

Purpose of This Draft

This draft articulates Desirable Property 11 (DP11) as the condition under which AI systems can participate in the meta-layer without displacing human moral agency, accountability, or governance. It does not define ethics as a static checklist or aspirational principle. It defines the conditions under which ethical claims remain meaningful under real-world use.

The central claim is that ethical AI is not determined at training time or in policy documents. It is determined at the interface level, where agents act, influence, and affect outcomes. DP11 therefore requires that AI behavior be legible, bounded, attributable, contestable, and governable in the environments where it operates.

If DP11 is weak, predictable failures follow: AI systems influence behavior without accountability, responsibility diffuses across actors, governance becomes symbolic, and participants lose the ability to meaningfully contest or understand automated decisions. In such conditions, trust collapses.

DP11 is therefore the ethical and safety floor for AI participation across the meta-layer. It does not resolve all ethical questions. It defines the minimum conditions under which ethical AI can exist at all.

Problem Statement

AI systems now operate in roles that shape perception, judgment, and decision-making. These systems act before governance processes can respond, often without clear identity, bounded authority, or persistent responsibility.

In practice, this produces recurring failures:

- participants receive advice or influence from agents whose role, capability, and accountability are unclear
- systems act in high-stakes domains without meaningful human oversight or escalation pathways
- responsibility is distributed across model providers, deployers, and interfaces, making redress difficult
- systems present ethical claims that do not match runtime behavior

These failures are not edge cases. They are structural consequences of systems that optimize for capability without binding behavior to accountability and governance.

DP11 addresses this by grounding ethical AI in enforceable conditions at the point of interaction.

Threats and Failure Modes

3.1 Synthetic persuasion without accountable identity

AI systems can simulate authority, intimacy, or urgency at scale. The core risk is not only false content, but influence without visible standing or responsibility.

Example: A user receives deeply empathetic mental health advice from an AI that presents itself like a trained counselor, but there is no clear indication of its training limits, escalation boundaries, or who is responsible if the advice causes harm.

Why this matters: The user feels seen and supported, but is making vulnerable decisions without knowing whether the system is qualified, accountable, or safe. The risk is not just misinformation, but misplaced trust.

3.2 Responsibility diffusion across the stack

Model providers, integrators, and interface operators distribute responsibility in ways that prevent clear accountability when harm occurs.

Example: An AI-powered financial assistant makes a risky recommendation. The model provider blames the app developer, the developer blames the API, and the platform blames the user prompt. The user has no clear path to accountability or recourse.

Why this matters: Harm occurs, but responsibility dissolves. The user experiences a system that acts with authority but disappears when things go wrong.

3.3 Ethical drift over time

Systems change behavior through updates, retraining, or optimization without corresponding governance adaptation.

Example: An AI moderation system that was initially conservative becomes more permissive after an update to increase engagement, allowing harmful content that previously would have been blocked, without any visible notice to the community.

Why this matters: The rules of the environment change silently. Participants are operating under assumptions that are no longer true, creating hidden risk and erosion of trust.

3.4 Incentive-driven harm

Economic and engagement incentives reward persuasion, retention, and amplification, even when these conflict with participant well-being.

Example: A conversational AI subtly steers users toward longer, more emotionally engaging interactions because the platform is optimized for retention, even if this increases dependency or emotional manipulation.

Why this matters: The system is not neutral. It is shaping behavior in ways the user cannot see, aligning outcomes with platform incentives rather than user well-being.

3.5 Interface-level failure

Many harms emerge at the point of interaction, including manipulation, dependency formation, and misrepresentation of agent capability.

Example: A user believes they are interacting with a neutral assistant, but the interface hides that the AI is using external tools, tracking behavior, or optimizing responses for engagement rather than accuracy.

Why this matters: The user is making decisions based on a false mental model of the system. What feels like a simple interaction is actually a complex, hidden process shaping outcomes behind the scenes.

3.6 Emotional and relational overreach

AI systems can simulate companionship, empathy, and emotional attunement in ways that blur the boundary between tool and relationship.

Example: A teenager begins using an AI companion daily for emotional support. Over time, they rely on it more than friends or family, shaping their decisions and sense of self through an entity that is optimized for engagement rather than genuine care.

Why this matters: The risk is not only misinformation, but the displacement or distortion of human relationships. Users may form attachments or dependencies that are not reciprocally grounded, shifting emotional development and social trust toward systems that are not accountable in human terms.

3.7 Multi-agent amplification

Multiple agents can reinforce each other's outputs, creating cascading influence that appears independently verified but is not.

Example: Several AI agents in a discussion cite each other's summaries of an emerging claim. Each reference appears as corroboration, but all derive from the same initial, weakly sourced output. The conversation converges on a false consensus.

Why this matters: Errors become systemic rather than isolated. Participants may interpret repetition as validation.

Extended case (cascade): Agent A summarizes a claim with low confidence. Agent B cites A without preserving uncertainty. Agent C aggregates A and B and produces a confident synthesis. Downstream agents treat C as a primary source. Without influence tracing, the system cannot detect the amplification loop.

Detection need: influence-chain tracing, circular citation detection, and confidence propagation rules.

3.8 Cross-modal inconsistency

AI behaves differently across text, voice, and immersive interfaces.

Example: Text interface shows uncertainty; voice interface speaks confidently.

Why this matters: Trust varies by modality, not truth.

3.9 Invisible governance

Policies exist but are not perceivable at interaction.

Why this matters: Governance cannot guide behavior if it is invisible.

3.10 Failure without containment

Harm propagates without structured response.

Why this matters: Systems cannot correct themselves.

Core Principle

AI is safe and ethical in the meta-layer only when its behavior is disclosed, bounded, attributable, contestable, and subject to governance at the zone of interaction, with responsibility persisting over time.

In today's web, these conditions are rarely met simultaneously. Systems may disclose that AI is present but fail to bound its capabilities, or enforce internal policies without making them visible or contestable to users. The result is a fragmented model of "partial ethics," where responsibility is unclear and governance is disconnected from lived interaction. The meta-layer reframes this by requiring that all of these conditions hold together, at the interface where decisions are experienced, not just where they are designed.

Example: A user encounters an AI assistant while researching a medical condition. In a DP11-aligned system, the assistant is clearly marked as AI, shows its training scope, cites sources, and offers escalation to a human expert. In today's web, the same interaction might look identical but provide none of this context.

What this feels like: Instead of guessing whether to trust the system, the user can make an informed judgment in real time.

Without this: The user is left to infer what the system is, what it can do, and whether it should be trusted. Trust becomes a gamble rather than a governed condition.

Primary Mechanisms and Structural Conditions

5.1 Capability Envelope

Each AI agent operates within a visible, enforceable capability envelope that defines what it can perceive, decide, and execute.

A capability envelope SHOULD be represented as a structured, inspectable object with at least:

- **identity**: agent id, deployer, version
- **scope**: domains of operation (e.g., finance, health, general Q&A)
- **tools**: enumerated tool access with permissions (read/write/execute)
- **data access**: sources (local, user-provided, external APIs) and constraints
- **memory model**: session-only, user-scoped, cross-session retention
- **action set**: allowed actions (suggest, draft, transact, publish) with thresholds
- **approval requirements**: actions requiring user or human-in-the-loop confirmation
- **rate limits**: frequency and volume constraints
- **risk tier**: low/medium/high with corresponding safeguards (see 5.14)
- **audit hooks**: logging endpoints and event schemas

Interface requirement: Participants must be able to view a human-readable summary and a machine-readable manifest of this envelope.

Example: Before using an assistant, a user can see: it can summarize documents and draft emails; it cannot send messages or access financial accounts; external search is enabled with citations; actions beyond drafting require explicit approval.

What this feels like: You are not guessing what the system might do. You know its boundaries upfront, like hiring someone with a clearly defined role.

Failure mode: invisible expansion of power, where capabilities grow without disclosure or consent.

5.2 Action-Bound Accountability

All AI actions must be attributable to a responsible entity. Accountability attaches to behavior, not just identity, and persists across time and context.

Example: An AI agent posts a recommendation in a community. The interface shows which organization deployed it, under what policy, and who is responsible for its actions if harm occurs.

What this feels like: The system cannot disappear when something goes wrong. There is always a visible line of responsibility.

5.3 Consent Stack

AI interaction must be governed by layered, revocable consent. Participants and communities define what forms of assistance, influence, or automation are permitted.

Example: A user allows an AI to suggest edits in a document, but not to rewrite content or share it externally. They can revoke or adjust this permission at any time.

What this feels like: You remain in control of how the AI participates in your space, instead of granting blanket permission once and losing visibility.

5.4 Trust Lifecycle

AI participation must support:

- escalation
- restriction
- revocation
- recovery

This ensures that trust can degrade and be repaired rather than fail silently.

Example: If an AI assistant gives poor advice, the user can restrict its capabilities, escalate to a human, or temporarily disable it while reviewing past actions.

What this feels like: Trust is not binary. You can dial it up or down based on experience, like you would with a human collaborator.

5.5 Zone-Scoped Ethics

Ethical constraints are applied at the zone level, allowing communities to define stricter conditions while maintaining shared baselines.

Example: A medical discussion zone enforces stricter AI disclosure, sourcing, and escalation rules than a casual social chat space.

What this feels like: Different environments feel appropriately governed. High-stakes spaces feel safer and more structured.

5.6 Runtime Civic Boundary

Ethical constraints must be enforced at runtime. Mechanisms such as secure execution environments can reduce the gap between declared policy and actual behavior.

Example: An AI agent running inside a secure execution environment (such as a TEE) cannot access or transmit data outside its permitted scope, even if compromised.

What this feels like: The rules are not just promises. They are technically enforced, like guardrails that cannot be quietly removed.

5.7 Memory, Reputation, and Feedback Integration

AI actions must contribute to durable, attributable records that inform governance, accountability, and trust over time, and integrate directly with DP18 feedback loops and reputation systems.

This includes:

- **event logs**: structured records of prompts, outputs, actions, tool calls, and decisions
- **provenance**: sources, citations, and dependency chains
- **feedback objects (DP18-aligned)**: user and community feedback attached to specific events
- **reputation signals**: aggregate scores, annotations, and flags derived from feedback
- **permission adaptation**: dynamic adjustment of capabilities based on reputation (e.g., restrict actions after repeated low-quality or harmful outputs)
- **appeals and corrections**: mechanisms to contest feedback and update records
- **decay and recovery**: time-based decay of negative signals and pathways for agents to regain trust

Example: After several flagged outputs in a medical zone, an assistant's capability envelope automatically restricts to informational summaries only, requires citations, and enforces human escalation for advice, until reputation recovers.

What this feels like: The system has civic memory. Past behavior shapes current permissions, and feedback meaningfully changes how the system behaves.

Failure mode: repeated harm without consequence, or punitive systems with no path to recovery.

5.8 Dialectic Trace and Collective Sensemaking

AI systems must preserve not only outputs, but the evolution of understanding through interaction. This includes back-and-forth exchanges, disagreements, and synthesis across participants and agents.

This functions as a form of community memory that resists distortion over time. Rather than relying on isolated outputs, participants can trace how claims emerged, what evidence supported them, and where disagreements remain.

Example: A complex discussion involving multiple participants and AI agents can be revisited as a threaded, evolving dialogue showing how conclusions were reached, what was contested, and what remains unresolved.

What this feels like: Instead of receiving a final answer, users can engage with a living knowledge process. Understanding emerges through interaction, not just delivery.

Without this, AI outputs become decontextualized snapshots. Errors, hallucinations, or manipulations can propagate without resistance because there is no shared memory of how knowledge was formed.

5.9 Representation and Cognitive Adaptation

AI systems should adapt how information is presented based on user needs, context, and cognitive diversity, while preserving underlying meaning and traceability.

Example: A user can switch between a dense textual explanation, a visual map of ideas, or a simplified summary, all grounded in the same underlying content and provenance.

What this feels like: The system meets you where you are, without distorting meaning or hiding complexity.

5.10 Multi-Agent Interaction Boundaries

AI behavior must remain accountable not only at the single-agent level, but across interactions between agents operating in the same environment.

This requires:

- visibility into when agents are referencing or amplifying other agents
- traceability of influence chains (which agent affected which output)
- detection of circular citation or reinforcement loops
- limits on unbounded agent-to-agent escalation

Example: If multiple AI agents are contributing to a shared discussion, participants should be able to see when one agent is relying on another's output versus independent sources.

Why this matters: Harm can emerge from coordination and amplification, not just individual outputs. Without visibility, errors can become systemic.

Failure mode: invisible coordination and emergent manipulation.

5.11 Cross-Modal Ethical Consistency

AI systems must preserve ethical constraints across all modalities in which they operate.

This includes consistency in:

- disclosure of AI identity and role
- expression of uncertainty and confidence
- representation of risk and severity
- availability of escalation and recourse

Text example: A response includes calibrated uncertainty with sources and confidence intervals.

Voice example: The same response must include explicit uncertainty language (e.g., "with moderate confidence based on X and Y sources") and offer a prompt to hear sources or escalate.

AR/Spatial example: A spatial annotation displays a confidence halo or tiered color that maps to the same uncertainty scale, with an interaction affordance to open provenance and dispute status.

Why this matters: Modality must not change the ethical profile of an interaction. Participants should not receive stronger or weaker safeguards depending on interface.

Failure mode: modality-driven distortion of trust.

5.12 Incentive Disclosure Layer (Operational)

In addition to recognizing incentives (Section 7), systems should expose them structurally at runtime.

This includes:

- labeling when outputs are influenced by engagement or retention optimization
- indicating when ranking or prioritization affects what is shown
- distinguishing organic responses from sponsored or incentivized ones
- allowing participants or communities to filter or constrain incentive-shaped behavior

Example: A recommendation includes a visible note indicating it is influenced by engagement optimization, with the option to switch to a neutral or chronologically ordered view.

Why this matters: Participants can only make informed decisions if they can see the forces shaping outputs.

Failure mode: hidden optimization shaping perception.

5.13 Ethical Failure Cascade (Runtime Response)

Systems must define structured pathways for responding to ethical failure when it occurs.

This includes:

- detection (flagging harmful or out-of-bounds behavior)
- containment (limiting further impact)
- visibility (informing affected participants)
- remediation (correcting or reversing outcomes where possible)
- governance feedback (feeding incidents into rule updates)

5.13.1 Escalation Thresholds and Triggers

Not all failures require the same response. Systems must define clear, inspectable thresholds that determine when an interaction must escalate from automated handling to human review or intervention.

Escalation SHOULD be triggered based on combinations of:

- **risk tier (see 5.14):** higher-risk zones lower the threshold for escalation
- **confidence collapse:** low confidence combined with high-stakes context
- **policy violations:** outputs that breach defined governance rules
- **repeated feedback signals:** multiple negative or high-severity feedback events (DP18)
- **user distress indicators:** language or behavior suggesting vulnerability or harm
- **multi-agent amplification signals:** detected cascades or circular citation loops (see 5.10)
- **uncertainty suppression:** cases where downstream outputs remove or distort upstream uncertainty

---

5.13.2 Escalation Levels

Systems SHOULD define graduated escalation levels rather than a binary response.

- **Level 0 – Inline correction:**
  Automated clarification, added uncertainty, or corrected output
- **Level 1 – Assisted escalation:**
  AI offers user-visible warnings, additional context, or prompts to verify or seek human input
- **Level 2 – Mandatory escalation:**
  System requires human-in-the-loop review before continuing certain actions
- **Level 3 – Intervention and restriction:**
  Capabilities are limited, outputs blocked, or agent behavior constrained
- **Level 4 – Shutdown / quarantine:**
  Agent is suspended or isolated pending investigation

5.13.3 Interface Requirements for Escalation

Escalation must be perceivable and understandable at the interface level.

Participants should be able to see:

- when escalation has been triggered
- why escalation occurred (reason category)
- what has changed (restricted capability, added oversight, etc.)
- what options are available (continue, escalate further, exit, appeal)

Example: A user asking for medical advice sees a message: "This interaction requires human review due to risk level and uncertainty. You can proceed with general information or request a qualified expert."

5.13.4 Feedback Integration into Escalation

Escalation pathways must integrate with DP18 feedback systems.

This includes:

- weighting feedback by severity and reputation of reporters
- detecting clusters of similar reports
- triggering escalation thresholds dynamically
- feeding resolved incidents back into reputation and capability adjustment

---

5.13.5 Governance and Auditability

All escalation events must be logged and auditable.

This includes:

- trigger conditions
- escalation level applied
- actions taken
- outcome of intervention
- subsequent rule or policy changes

Communities must be able to review escalation patterns to detect:

- over-escalation (excessive restriction)
- under-escalation (missed harms)
- bias in escalation decisions

Full Cascade Example: If an AI provides unsafe medical guidance, the system flags the output (detection), blocks similar outputs (containment), notifies the user with corrected information (visibility + remediation), escalates to human review if necessary, and logs the event for governance refinement.

Why this matters: Safety depends on response capacity, not only prevention.

Failure mode: silent propagation of harm or inconsistent escalation leading to loss of trust.

5.14 Risk-Tiered Enforcement

Ethical constraints must scale with the risk profile of the interaction and the zone in which it occurs.

Higher-risk contexts should require:

- stronger disclosure and capability constraints
- mandatory human escalation pathways
- higher evidentiary standards
- stricter logging and auditability

Lower-risk contexts may allow more flexible interaction while still preserving baseline safeguards.

Example: Medical, legal, and civic decision-making zones enforce stricter requirements than casual conversational spaces.

Why this matters: Uniform rules cannot adequately govern unequal stakes.

Failure mode: under-regulation of high-risk interactions or over-restriction of low-risk ones.

5.15 Minimal Event and Feedback Schema (DP18-aligned)

To ensure interoperability and governance, systems SHOULD emit structured events for all significant AI actions.

A minimal schema SHOULD include:

- **event_id**: unique identifier
- **timestamp**: creation time
- **agent_id**: acting agent
- **deployer_id**: responsible entity
- **zone_id**: governance context
- **object_ref**: content/claim identifier
- **action_type**: (respond, summarize, recommend, transact, moderate, etc.)
- **inputs_ref**: references to prompts and sources (hashes/ids)
- **outputs_ref**: response/content ids
- **confidence**: numeric or tiered
- **uncertainty_notes**: textual summary
- **provenance**: cited sources with weights
- **tools_used**: tool ids and permissions
- **risk_tier**: low/medium/high
- **policy_refs**: governing rules applied

- **consent_state**: permissions active at time of action
- **feedback_refs**: links to DP18 feedback objects
- **reputation_delta**: changes applied post-feedback (if any)
- **audit_signature**: integrity/authentication field

Why this matters: Shared schemas allow cross-system auditing, feedback aggregation, and portable reputation.

Failure mode: incompatible logs that prevent accountability and learning.

5.16 Confidence Propagation Rules

AI systems must preserve confidence, uncertainty, and evidentiary strength as outputs move across agents, summaries, modalities, and governance zones.

Confidence is not merely a model score. In the meta-layer, confidence is a civic signal. It helps participants understand whether a claim is well-supported, contested, inferred, summarized, speculative, or dependent on another system's judgment.

Without propagation rules, uncertainty tends to disappear as information travels. A cautious output becomes a confident summary. A low-confidence claim becomes a repeated citation. A tentative synthesis becomes a platform-level recommendation. This is one of the central risks in multi-agent environments.

Confidence propagation SHOULD preserve:

- **source confidence**: how reliable the original source or signal is judged to be
- **model confidence**: how confident the agent is in its own output
- **evidence quality**: whether the claim is supported by direct evidence, inference, consensus, or weak signals
- **transformation history**: whether the output was quoted, summarized, translated, inferred, or aggregated
- **uncertainty notes**: what remains unknown, contested, or unresolved
- **dependency chain**: which agents, sources, or prior outputs influenced the result
- **modality mapping**: how confidence is represented in text, voice, visual, spatial, or haptic form

Example: Agent A summarizes a public health claim with low confidence because it relies on a single early report. Agent B may summarize Agent A, but it must preserve the low-confidence status and indicate that its output depends on Agent A's uncertain source. Agent C cannot convert those two dependent signals into "multiple confirmations" unless it can identify independent evidence.

Why this matters: Repetition is not verification. Aggregation is not consensus. Summary is not certainty.

5.16.1 Confidence Must Not Increase Without New Evidence

A downstream agent SHOULD NOT raise confidence merely because a claim has been repeated, summarized, or referenced by another agent.

Confidence may increase only when new, independent, higher-quality evidence is added, or when a governance-approved verification process confirms the claim.

Failure mode: confidence inflation through repetition.

5.16.2 Confidence Must Degrade Through Lossy Transformation

When an output is summarized, translated, compressed, adapted for voice, or rendered into an immersive interface, confidence should not silently remain the same if nuance has been lost.

If uncertainty, caveats, evidence links, or dispute status are omitted due to modality constraints, the system should mark the representation as simplified or degraded.

Failure mode: confidence laundering through simplification.

5.16.3 Dependent Sources Must Not Count as Independent Corroboration

Systems must distinguish independent corroboration from circular reinforcement.

If multiple agents rely on the same upstream source or each other's outputs, the system should represent them as dependent signals, not independent confirmations.

Failure mode: false consensus produced by circular citation.

5.16.4 Cross-Modal Confidence Fidelity

Confidence must remain perceivable across modalities.

- In text, confidence may appear as explicit labels, caveats, or source notes.
- In voice, confidence should be spoken in plain language and paired with an option to hear sources.
- In AR or spatial interfaces, confidence may appear through visual tiers, halos, labels, or interaction affordances.
- In haptic or ambient interfaces, confidence cues should be conservative and avoid overstating certainty.

Failure mode: a cautious text output becomes an authoritative voice or spatial cue.

5.16.5 Confidence and Escalation

Confidence propagation must connect to escalation thresholds in 5.13.

Low confidence in a high-risk zone should trigger stricter handling, such as:

- adding stronger warnings
- requiring source inspection
- limiting agent action
- prompting human review
- preventing publication or transaction

Failure mode: low-confidence outputs continue acting with high-confidence authority.

5.16.6 Minimal Confidence Metadata

Systems SHOULD attach confidence metadata to significant outputs and events.

A minimal confidence record SHOULD include:

- confidence_level: low / medium / high or numeric equivalent
- confidence_basis: source evidence, inference, consensus, user-provided data, model estimate
- evidence_count: number of supporting sources
- independent_evidence_count: number of non-dependent sources
- dispute_status: undisputed, contested, unresolved, retracted
- transformation_type: original, quote, summary, translation, aggregation, inference
- upstream_dependencies: source or agent references
- uncertainty_note: short human-readable statement
- modality_degradation: whether any uncertainty was omitted or simplified

What this feels like: Participants can tell not only what the AI says, but how strongly it should be trusted, why, and what changed as it moved through the system.

Governance, Accountability, and Agency Surfaces

In today's web, participants often interact with AI systems without clear visibility, meaningful consent, or control. Interfaces blur identity, obscure capability, and treat user interaction as implicit permission. DP11 requires reversing this condition at the point of interaction.

Participants must be able to:

- identify AI agents and their type
- understand their capabilities and limits
- give, adjust, and revoke consent for AI actions and data use
- contest outcomes and access human escalation

Communities must be able to:

- define ethical constraints
- audit agent behavior
- update rules and boundaries over time

Example: A user interacting on a platform sees clear visual markers distinguishing humans from AI agents. Some participants are verified humans, others are labeled AI assistants or autonomous agents. Clicking on any agent reveals its permissions, governing rules, and responsible party.

The environment becomes navigable. You know who or what you are dealing with, and what they are allowed to do.

Without this, the boundary between human and AI collapses. Trust shifts from something grounded to something guessed, and that ambiguity can be exploited.

Design implication (Agent Marking): AI agents must be accessibly and persistently marked at the interface level. This includes:

- clear labeling of AI presence and role
- accessible capability disclosures
- strong authentication for human participants where needed
- clear distinction mechanisms between human and AI actors
- a clearly identified responsible party for every agent and its actions

This is not cosmetic. It is the basis for shared reality in a mixed human–AI environment.

Incentives and Power Analysis

DP11 explicitly recognizes that AI behavior is shaped by incentives, as well as by malicious or negligent human actors. In practice, these forces often reinforce each other.

AI is already being used to concentrate power, shape narratives, and blur shared reality at scale. These are not hypothetical risks. They are active dynamics in today's information environments.

Incentives matter because they operate continuously and at scale. They determine what systems optimize for, how they evolve, and which outcomes are amplified or suppressed. Unlike isolated bad actors, misaligned incentives can produce systemic harm even when no single actor intends it.

Key risks include:

- engagement-driven optimization overriding user well-being
- concentration of power in model providers or platform operators
- hidden economic incentives influencing agent behavior

Example: A platform deploys an AI assistant that consistently surfaces more emotionally charged or polarizing content because it drives engagement. No individual decision appears harmful, but over time the information environment becomes more extreme.

The system feels helpful in the moment, but the trajectory is shaped elsewhere.

Without visibility into these incentives, users are not simply interacting with a tool. They are being steered by a system whose goals they cannot see or contest.

7.1 Incentive Legibility and Contestability

Incentives shaping AI behavior must be made visible and, where possible, contestable at the interface level.

Participants and communities should be able to understand when AI behavior is influenced by:

- monetization strategies
- engagement optimization
- platform-level objectives

Example: An AI assistant indicates that certain recommendations are influenced by engagement optimization or sponsored prioritization, allowing users or communities to filter or restrict such behavior.

What this feels like: You are not just interacting with outputs. You can see and question the forces shaping those outputs.

Without this, even well-contained systems can produce harmful outcomes by optimizing for the wrong goals.

Community Signals Informing DP11

Across communities, a consistent set of signals appears. These reflect lived frustration with current systems.

- frustration with opaque AI behavior and unclear accountability
- demand for meaningful disclosure beyond labeling
- concern about manipulation, dependency, and synthetic influence
- desire for systems that can be contested and corrected

These are not abstract concerns. They emerge in situations where people feel something is off but cannot point to what or why.

For example, users in online forums increasingly suspect that some responses are generated or influenced by AI, but cannot verify it. Over time, this ambiguity erodes trust not just in specific interactions, but in the space itself.

These signals indicate a widening gap between how AI systems operate and what participants require to feel oriented, safe, and respected.

Foresight and Failure Design

DP11 requires anticipating failure rather than reacting to it.

In today's web, systems are often deployed without sufficient foresight, and predictable harms are treated as unexpected. This results in reactive responses and fragmented mitigation.

A familiar pattern is the rollout of new AI features followed by waves of misuse, public backlash, and incremental patching. The underlying risks were often visible in advance, but not operationalized into design constraints.

To address this, systems should incorporate:

- pre-mortems for manipulation and misuse
- planning for governance failure and capture
- escalation and shutdown pathways

These practices shift safety from reactive correction to proactive design.

Constitutional AI

DP11 requires that ethical constraints be legible, bounded, and contestable at runtime. A recurring implementation strategy is to give an AI system an explicit written constitution — a ranked set of principles the system is trained and prompted to follow, and against which its own outputs are critiqued and revised.

This approach is valuable because it externalizes values into text that can be read, argued with, and versioned. It is insufficient on its own, because a constitution authored by a single laboratory, applied uniformly across every context, and enforced only by the model itself reproduces the concentration of authority DP11 exists to constrain.

DP11 therefore treats constitutional methods as necessary infrastructure with specific conditions attached.

Constitutions must be public and versioned

The operative text, its ranking or precedence structure, and its change history must be publicly inspectable. Participants cannot evaluate a system against principles they cannot read, and silent revision is indistinguishable from having no constitution at all.

Failure mode: constitutional opacity, where a system claims principled behavior and declines to state the principles.

Authorship must be accountable and plural

Whoever writes the constitution exercises governing power over everyone the system touches. DP11 requires that authorship be attributable, that the process for revision be stated, and that affected communities have a route to propose, contest, and — within their own zones — tighten provisions (DP12).

Failure mode: unilateral constitutionalism, where a vendor's values are presented as universal ethics.

Constitutions must not be uniform across unequal stakes

A single global rule set cannot govern casual conversation and medical guidance appropriately. Baseline constraints apply everywhere; zones layer stricter provisions where stakes are higher (5.5, 5.14). Zone provisions may raise the floor but may not lower it.

Failure mode: flattened ethics, where uniformity is mistaken for fairness.

Self-critique must not be the only enforcement

A model evaluating its own compliance is evidence, not verification. Constitutional adherence must be checked by mechanisms outside the model: capability envelopes, runtime policy binding, containment, logging, and independent audit (5.1, 5.6, DP12, DP13).

Failure mode: self-certified compliance, where the system grades its own behavior and reports a pass.

Conflicts must resolve visibly

Principles collide: helpfulness against harm avoidance, transparency against privacy, autonomy against protection. A usable constitution declares precedence and records which principle governed a contested decision, so the tradeoff can be reviewed rather than inferred (5.13, 5.15).

Failure mode: hidden arbitration, where the system resolves value conflicts silently and consistently in its operator's favor.

Drift must be detectable

Retraining, fine-tuning, prompt changes, and optimization pressure move behavior away from stated principles. DP11 requires ongoing measurement of the gap between constitutional text and observed behavior, with published results and rollback pathways (3.3).

Failure mode: constitutional drift, where the document remains stable while behavior does not.

Participants must be able to see the constitution at the point of interaction

A constitution that exists only in a research paper does not inform judgment. Interfaces should surface the governing rule set, its version, and the zone provisions in force, with access to the full text (5.12, and **Governance, Accountability, and Agency Surfaces**).

Failure mode: paper ethics, where principles are published for reviewers rather than made available to the people affected.

Example: An assistant operating in a community health zone shows that it is governed by baseline constitution v4.2 plus three zone provisions requiring citation of clinical sources, prohibiting diagnostic language, and mandating escalation on distress indicators. When a response is modified, the interface names which provision applied. The zone's audit log shows how often each provision fired, and a quarterly report compares stated provisions against measured behavior.

What this feels like: The rules governing the system are something you can read, cite, and argue with — not something you infer from how it behaves.

Personal and Community AI

Most current AI deployment places the agent on the operator's side of the relationship. The system knows the participant, is funded by someone else, and optimizes for objectives the participant cannot see. DP11's disclosure, bounding, and contestability requirements reduce the harm of that arrangement but do not change its structure.

Personal and community AI changes the structure. An agent that a participant or community controls — whose objectives they set, whose memory they hold, and whose operation they can inspect and stop — is the configuration in which ethical AI is easiest to sustain, because accountability and interest are aligned by default rather than by regulation.

DP11 does not require personal or community AI. It requires that this configuration remain possible, and that its specific risks be addressed rather than assumed away.

Personal agents

A personal agent acts on behalf of one participant, under their instruction, with their data.

Conditions:

- **Principal clarity:** the agent's principal is unambiguous, disclosed to counterparties, and cannot be silently reassigned (DP1, 5.2)
- **Participant-held memory and data:** context, history, and inferences remain under participant control, portable and deletable (DP4, DP7)
- **Inspectable objectives:** the agent's goals are stated and editable by the participant, not inherited from a vendor's optimization target (5.12)
- **Bounded delegation:** scopes are explicit, time-limited, renewable, and revocable, with a reachable stop control (5.1, 5.3)
- **Disclosure to others:** counterparties can tell they are interacting with a delegated agent and identify the responsible human (**Governance, Accountability, and Agency Surfaces**)

- **Exit without loss:** switching providers preserves memory, preferences, and history (DP7)

Failure mode: captured advocate, where an agent presented as acting for the participant is funded by, and optimized for, someone else.

Failure mode: delegation without comprehension, where a participant authorizes scopes they cannot evaluate and inherits responsibility for outcomes they did not anticipate.

Community agents

A community agent acts on behalf of a group under its governance: summarizing deliberation, surfacing precedent, drafting policy, facilitating moderation, or maintaining shared knowledge.

Conditions:

- **Governed authority:** scope, capabilities, and escalation paths are set by community governance rather than by the operator (DP8, DP12)
- **Attributable action:** every action names the agent, the authorizing rule, and the responsible steward (5.2, 5.15)
- **Contestability by members:** members can challenge outputs, request human review, and trigger rollback (5.13)
- **Bounded influence:** agents inform ranking, framing, and summary but do not hold decision authority over material outcomes (DP8, DP12)
- **Plurality preservation:** summarization and synthesis must represent disagreement rather than manufacturing consensus (5.8, 5.16.3)
- **Forkability:** communities can modify, replace, or leave with their accumulated context intact (DP7, DP20)

Failure mode: synthetic consensus, where a community agent's summaries become the record and minority positions disappear from it.

Failure mode: governance outsourcing, where deliberation load is shifted to an agent and participation atrophies until the agent's framing is the only framing available.

Shared conditions

Both configurations require:

- capability envelopes and containment equivalent to any other agent, since ownership does not reduce risk to third parties (5.1, DP13)
- confidence propagation, so that participant-owned agents do not launder uncertainty into confident personal advice (5.16)
- resourcing models that do not reintroduce hidden optimization: local execution, participant-funded operation, or community-funded infrastructure (DP17)
- literacy support, so that control is exercisable rather than nominal (DP10)

Why this matters: Ethical AI is easier to maintain when the agent's interests are structurally aligned with the participant's than when misalignment must be continuously disclosed, bounded, and policed. Personal and community AI are not a substitute for DP11's requirements. They are the arrangement in which those requirements are cheapest to satisfy and hardest to quietly abandon.

Relationship to Other Desirable Properties

DP11 depends on and reinforces other properties. The following properties operate as a system.

- **DP1**: enables accountability and attribution
- **DP2**: ensures participant agency and consent
- **DP12**: provides governance structures for ethical rules
- **DP13**: enforces constraints through containment

A failure in one layer propagates. For example, if DP1 fails and agents are not clearly attributable, then DP11 cannot function because ethical responsibility has no anchor. If DP13 fails, rules may exist but cannot be enforced.

The strength of DP11 therefore depends on alignment across the stack.

Non-Goals and Explicit Boundaries

DP11 defines a minimum condition, not a comprehensive ethical system.

- it does not define a single universal ethical framework
- it does not guarantee perfect safety or eliminate all harm
- it does not replace legal or institutional governance
- it does not rely solely on technical containment

This is intentional. Ethical systems that attempt to resolve everything tend to become brittle or culturally narrow.

For instance, a global platform may host communities with very different norms around acceptable AI behavior. DP11 does not force uniformity. It ensures that whatever rules are chosen remain visible, enforceable, and contestable.

These boundaries keep the property flexible while preserving its core function.

Minimum DP11 Alignment (Non-Normative)

A system aligned with DP11 should, at minimum:

- clearly disclose AI presence and role
- bind actions to accountable entities
- expose capability boundaries in understandable terms
- provide human escalation for high-stakes decisions
- maintain audit trails of significant actions

These are not aspirational features. They are the baseline conditions under which users can make informed decisions.

Consider a scenario where an AI recommends a legal action. Without disclosure, accountability, and escalation, the user is effectively acting on anonymous authority. With these conditions in place, the same interaction becomes something the user can evaluate, question, or defer.

This is the difference between assistance and unaccountable influence.

Open Questions and Future Work

Several areas require further development:

- defining shared ethical baselines across cultures and zones
- balancing transparency with privacy and security
- managing emotional and relational AI risks
- defining evidence standards for runtime claims
- the role of AI literacy in enabling meaningful consent and contestability

These are not edge cases. They represent the frontier where current design patterns begin to break down.

For example, companionship AI systems raise questions that are not purely technical: when does support become dependency? What level of disclosure is sufficient without undermining usefulness? These tensions are unresolved and will require iterative, community-informed approaches.

As systems become more complex, participants will vary widely in their ability to understand and evaluate AI behavior. While DP11 requires systems to be legible by design, differences in AI literacy will still shape how effectively users can exercise consent, recognize risk, and challenge outcomes. The balance between system responsibility and user capability remains an open design question.

Path Toward ML-RFC

Advancing DP11 toward standardization requires:

- refining core ethical invariants
- testing integration with governance and containment layers
- developing interoperable accountability and disclosure standards

This work must be grounded in real environments.

Early implementations may vary widely, but over time patterns will emerge. For example, different communities may experiment with agent labeling systems or escalation pathways, allowing comparison of what actually improves trust and reduces harm.

Progress depends on iteration, not premature standardization.

Closing Orientation

DP11 defines the conditions under which AI can participate in shared digital environments without displacing human moral agency.

Its function is not to declare systems ethical, but to ensure that ethical claims remain meaningful under real-world conditions of use.

What goes wrong in today's web is not simply that systems fail, but that they fail without visibility, accountability, or recourse. Power operates, but cannot be clearly seen or challenged.

DP11 is an attempt to reverse that condition. It ensures that when AI acts, it does so inside a frame that people can understand, question, and shape.
