DP14 – Trust and Transparency

Epistemic Integrity (V2) — the conditions under which participants can form reliable beliefs about system behavior

---

1. Purpose of This Draft

DP14 defines the conditions under which participants can form **reliable beliefs about system behavior**. It ensures that what users see, are told, and infer is **truthful, sufficiently complete, non-manipulative, and verifiably grounded**.

DP14 is the human-facing layer of DP15 (evidence & provenance). It binds interface signals to verifiable reality, connects to DP16 (truthful commitments), DP17 (incentives), DP8 (governance), DP4 (data), and DP12 (AI).

If DP14 fails, systems can be technically correct while **socially deceptive**, leading to miscoordination at scale.

2. Problem Statement

Modern systems routinely shape perception through:
- selective disclosure
- persuasive interfaces
- opaque algorithms
- AI-generated explanations

Participants cannot reliably determine:
- what is true vs. implied
- what is complete vs. omitted
- what is verified vs. asserted

This produces **epistemic drift**: beliefs diverge from underlying reality without detection.

DP14 reframes transparency as **epistemic integrity**: signals must be **truthful, checkable, and resistant to manipulation**.

3. Threats and Failure Modes (Adversarial Model)

DP14 assumes that adversaries will not only hide information. They will shape what participants believe through partial truths, plausible explanations, interface framing, and synthetic authority.

Transparency can itself become an attack surface when systems disclose enough to appear honest while withholding, reframing, or fabricating the context needed for understanding.

3.1 Selective transparency

Systems disclose true fragments while omitting context that would change interpretation.

Example: a ranking system reveals that "quality signals" affect visibility but omits that paid placement, engagement pressure, or internal partnership status dominates the outcome.

Failure mode: **truthful misdirection**, where disclosed facts are technically accurate but epistemically misleading.

3.2 Explainability theater

Systems generate explanations that sound plausible but are not grounded in actual decision pathways.

Example: an AI moderation system tells a participant that content was removed for "community safety" while the actual trigger was an automated keyword rule or advertiser exclusion list.

Failure mode: **synthetic explanation**, where explanation substitutes for accountability.

3.3 AI narrative shaping

AI systems produce fluent summaries, warnings, or justifications that overstate certainty, capability, neutrality, or consensus.

Example: an AI-generated summary presents a contested issue as settled by selectively compressing sources and omitting dissenting evidence.

Failure mode: **plausibility capture**, where fluency and confidence override uncertainty and evidence.

3.4 Interface manipulation

UI framing, ordering, visual weight, defaults, and timing shape interpretation without explicit falsehood.

Example: a "verified" badge is visually emphasized while the underlying verification only confirms payment or account control, not expertise or trustworthiness.

Failure mode: **perception steering**, where interface design causes participants to infer stronger claims than the system can support.

3.5 Trust signal spoofing

Bad actors simulate legitimacy through forged, contextless, or inflated indicators.

Example: coordinated accounts manufacture endorsements, badges, reputation scores, or "community consensus" signals to make content appear broadly trusted.

Failure mode: **synthetic legitimacy**, where trust indicators detach from accountable contribution or evidence.

3.6 Epistemic drift over time

Signals, explanations, or labels change without visible history, causing participants to lose track of what was previously represented as true.

Example: a platform silently revises the explanation for a recommendation, moderation decision, or AI output after challenge or criticism.

Failure mode: **memory erosion**, where belief history cannot be reconstructed.

3.7 Transparency overload

Systems disclose too much unstructured information, making meaningful understanding impossible.

Example: a participant is shown dozens of technical logs, model cards, policy references, and disclaimers without actionable synthesis.

Failure mode: **legibility collapse**, where disclosure volume defeats comprehension.

3.8 Strategic uncertainty laundering

Systems hide behind uncertainty even when they have enough information to disclose more precise risk or confidence levels.

Example: a system labels an output "AI-assisted" but refuses to distinguish between minor grammar support and full autonomous generation.

Failure mode: **ambiguity sheltering**, where uncertainty language protects operators from accountability.

3.9 Cross-system context loss

Transparency signals degrade as artifacts move across tools, platforms, zones, or interfaces.

Example: a provenance-backed warning appears in one overlay but disappears when the content is embedded elsewhere.

Failure mode: **context stripping**, where participants encounter content without the interpretive scaffolding needed to assess it.

3.10 Multi-vector epistemic attacks

Adversaries combine AI content, fake trust signals, selective evidence, interface timing, and coordinated amplification.

Example: a campaign uses AI-generated expert commentary, forged endorsements, plausible citations, and paid visibility to create the appearance of consensus.

Failure mode: **manufactured reality**, where multiple weak or manipulated signals reinforce one another into a false but convincing worldview.

4. Core Principle

> Participants must be able to form **accurate, bounded, and contestable beliefs** about system behavior.

This requires:
- **truthfulness** (no misleading representations)
- **bounded completeness** (enough context to avoid misinterpretation)
- **verifiability** (grounding in DP15 evidence)
- **contestability** (ability to challenge and correct)

Trust emerges from **reliable belief formation**, not persuasion.

5. Primary Mechanisms and Structural Conditions

5.1 Transparent environments

Visible context for rules, actors, and system state.
- MUST expose policy versions, decision mode (human/AI), and active constraints
- Verification: inspectable context + version history
- Failure: context opacity, hidden conditions

5.2 Algorithmic transparency

Explanations tied to actual decision logic.
- MUST distinguish local vs. global explanations and show uncertainty
- Verification: mapping to logs/provenance (DP15)
- Failure: explainability theater, inconsistency

5.3 AI transparency

Clear disclosure of model role, scope, and limits (DP12).
- MUST show model/version, capability class, and policy gates
- Verification: links to evals/attestations (DP15)
- Failure: capability misrepresentation, attribution ambiguity

5.4 Reputation systems

Explainable, provenance-backed trust signals.
- MUST show origin, criteria, scope, and decay
- Verification: trace to receipts/events (DP15)
- Failure: spoofing, opaque scoring

5.5 Behavioral standards

Legible rules with consistent enforcement.
- MUST link violations to actions and precedents
- Verification: audit logs (DP15)
- Failure: rule–enforcement mismatch, selectivity

5.6 Decision traceability

Lineage for governance and system decisions.
- MUST record rationale, inputs, actors, and versions (DP3)
- Verification: reconstruct decisions over time
- Failure: untraceable decisions, post-hoc rationalization

5.7 Transparency of incentives

Mapping from incentives to behavior (DP17, DP9).
- MUST disclose drivers (revenue, ranking, sponsorship)
- Verification: correlate with outcomes; audit preferential treatment
- Failure: hidden incentives, pay-to-play opacity

5.8 AI containment visibility

Visible policy boundaries and interventions.
- MUST indicate blocks, edits, uncertainty, escalation paths
- Verification: policy logs and consistency (DP15)
- Failure: invisible containment, boundary leakage

5.9 Real-time transparency signals

Timely indicators for state, risk, and verification.
- MUST show validity (valid/unknown/invalid) and uncertainty
- Verification: signals match underlying state
- Failure: signal suppression, overstated certainty

5.10 Cross-system transparency

Preservation across tools (DP7) with degradation signals.
- MUST use portable formats and indicate loss of context
- Verification: import/export consistency checks
- Failure: silent loss, incompatibility

5.11 Epistemic Integrity System Layer

Ensures signals remain **truthful, bound to evidence, and resistant to manipulation**.

5.11.1 Signal generation

Explanations tied to real policies, models, and decisions.
- Failure: synthetic transparency

5.11.2 Signal–evidence binding (DP15)

All claims trace to logs, artifacts, or attestations.
- Failure: unverifiable claims

5.11.3 Signal propagation

Preserved across systems with explicit degradation.
- Failure: transparency loss

5.11.4 Anti-deception constraints

Detect and mitigate selective disclosure, UI distortion, and AI misrepresentation.
- Failure: deceptive legibility

5.11.5 Contestability

Dispute, evidence request, and escalation pathways (DP3, DP8).
- Failure: non-contestable transparency

5.11.6 Trust signal integrity

Provenance-backed indicators; anti-Sybil protections.
- Failure: spoofed legitimacy

5.11.7 Memory and auditability

Versioned explanations and comparison over time.
- Failure: epistemic drift

6. Governance, Accountability, and Agency Surfaces

Participants MUST be able to:
- inspect signals and underlying evidence
- challenge misleading representations
- trigger reviews tied to specific items

Governance MUST be able to:
- require correction or reclassification of signals
- attach confidence ratings to transparency surfaces
- sanction repeated deception (downgrade, restrict features)

Failure modes:
- non-actionable transparency
- accountability gaps

7. Incentives and Power Analysis

Opacity and persuasion are often profitable.

Dynamics:
- attention/revenue tied to persuasive framing
- underinvestment in truthful explanation
- AI fluency used to overstate certainty

Attack surfaces:
- selective disclosure for advantage
- sponsored influence shaping visibility
- narrative manipulation via AI

Mitigations:
- disclose incentive structures alongside signals
- penalize repeated misleading transparency
- require evidence binding for high-impact claims

Failure modes:
- incentive inversion (misleading signals are rewarded)

8. Community Signals Informing DP14

Signals indicate demand for:
- explainable AI
- visible moderation and rules
- trustworthy indicators

Operationalization:
- metrics for explanation quality and consistency
- thresholds triggering review (e.g., mismatch rates)

Failure:
- signal neglect, performative transparency

9. Evaluation Criteria

DP14 is testable only if epistemic integrity is measured by **outcomes** rather than by the presence of disclosure surfaces. A system with extensive transparency UI can still score poorly if participants routinely form false beliefs from it.

9.1 Fidelity measures

- **explanation–decision match rate:** share of explanations that survive counterfactual or ablation testing
- **consistency rate:** similar inputs produce materially similar explanations
- **evidence binding coverage:** share of high-impact claims linked to retrievable DP15 artifacts
- **uncertainty accuracy:** stated confidence tracks observed correctness

9.2 Comprehension measures

- **task-based comprehension:** participants can correctly state what happened, why, and how certain it is
- **misinterpretation rate:** frequency of confident but incorrect inferences drawn from a surface
- **overload indicators:** abandonment, non-use, or skimming of disclosure surfaces
- **accessibility parity:** comprehension holds across languages, literacy levels, and assistive technologies

9.3 Contestability measures

- **dispute intake latency:** time from report to acknowledgment with a tracked identifier
- **correction latency:** time from valid dispute to published correction with diff
- **overturn rate:** share of disputes resulting in changed signals, explanations, or classifications
- **repeat-dispute rate:** recurrence of the same defect after correction

9.4 Durability measures

- **signal degradation rate across hops:** loss of context when artifacts move between systems (DP7)
- **history completeness:** share of signals with reconstructable version history
- **drift detection time:** interval between a signal diverging from reality and detection
- **spoofing detection rate:** identification of forged or inflated trust indicators

9.5 Thresholds and gating

Measurement must connect to consequences, or it becomes another transparency surface.

- high-risk zones carry higher fidelity and evidence-binding floors
- exceeding mismatch or misinterpretation thresholds triggers review (Section 6)
- repeated failures downgrade confidence ratings or restrict features that depend on the affected signals
- metrics have named owners, publication cadence, and methodology disclosure

Failure modes:

- **metric theater**, where favorable indicators are published while defects persist
- **survivorship measurement**, where only participants who persisted are measured
- **unowned metrics**, where results exist but no actor is responsible for acting on them

10. Non-Goals and Explicit Boundaries

DP14 does not require full disclosure of all internals.

It explicitly disallows:
- explainability theater
- selective disclosure that misleads
- UI patterns that distort meaning
- trust signals without provenance

Principle:

> Systems may simplify, but must not mislead.

11. Minimum Alignment (Non-Normative)

A DP14-aligned system MUST:
- bind explanations to verifiable evidence (DP15)
- provide sufficient context to avoid misinterpretation
- preserve signals across systems with degradation notices
- enable contestability and correction
- maintain history of signals and explanations

Failure modes:
- deceptive legibility
- unverifiable claims
- silent changes over time

Systems lacking evidence binding, contestability, or memory SHOULD NOT be considered aligned.

12. Open Questions and Future Work

DP14 requires operational answers to questions that determine whether transparency produces **reliable understanding** rather than noise or manipulation.

12.1 Explanation fidelity (provable faithfulness)

Define measurable standards for whether an explanation is **causally linked** to the decision pathway.

- Methods: counterfactual tests, feature ablations, rule tracing, policy matching
- Requirement: explanations MUST fail when the underlying decision changes
- Open problem: standardizing faithfulness across models (symbolic, statistical, hybrid)

12.2 Bounded completeness (anti-misleading thresholds)

Determine the **minimum context set** required to avoid misleading users.

- Define "misleading by omission" thresholds per use case
- Tiered disclosure: summary → details → raw evidence
- Role-based views: participant, auditor, steward

12.3 Transparency vs. security (safe disclosure envelopes)

Formalize disclosure envelopes that prevent:
- leaking exploit thresholds
- enabling evasion of safeguards

while still exposing:
- policy classes
- decision categories
- uncertainty and limits

12.4 AI explanation standards (self vs. external explanation)

Separate:
- **system-generated explanations** (prone to self-justification)
- **independent verification layers** (overlay auditors)

Define when external corroboration is required for high-impact decisions.

12.5 Cross-system preservation (loss models)

Specify loss models for transparency signals across systems:
- what fields must persist
- what degradation is acceptable
- how to signal loss to users

12.6 Measuring epistemic reliability (outcomes, not intent)

Define metrics such as:
- explanation consistency rate
- dispute overturn rate
- correction latency
- signal degradation rate across hops
- user comprehension accuracy (task-based)

12.7 Governance of transparency (who sets the bar)

Define:
- who can raise transparency requirements in high-risk zones
- how disputes over "misleading" are adjudicated
- escalation from local to cross-system governance

---

13. Relationship to Other Desirable Properties (Operational Binding)

DP14 converts other DPs into **perceivable and actionable reality**.

- **DP15 (Evidence):** DP15 provides proofs; DP14 defines how proofs are surfaced, summarized, and validated by users. Missing DP14 → evidence exists but is unusable.
- **DP16 (Commitments):** Roadmap claims must be presented with uncertainty, funding state, and change history. Missing DP14 → commitments appear firmer than they are.
- **DP17 (Finance):** Incentive disclosures must be legible and tied to behavior. Missing DP14 → hidden extraction persists behind complex reporting.
- **DP8 (Governance):** Decisions require visible rationale and contest paths. Missing DP14 → governance legitimacy degrades.
- **DP12 (AI):** Model scope, limits, and policy must be visible at interaction time. Missing DP14 → AI over-claim and misinterpretation.
- **DP4 (Data):** Collection, inference, and sharing must be explained at the point of impact. Missing DP14 → consent is uninformed.
- **DP20 (Ownership):** Rights and surplus flows must be legible. Missing DP14 → "ownership theater".

Constraint:

> No DP can claim alignment if its guarantees are not **legible, bounded, and contestable** at the interface.

---

14. Foresight and Failure Design (Epistemic Incident Model)

Treat transparency failures as **epistemic incidents** with lifecycle management.

Incident classes

- E1: Misleading explanation (unfaithful)
- E2: Omission-induced misinterpretation
- E3: Signal spoofing / fake trust indicators
- E4: Context loss across systems
- E5: AI overstatement / hallucinated justification

Detection

- anomaly detection on explanation–outcome mismatch
- user reports with reproducible cases
- cross-system inconsistency checks

Containment

- flag affected signals as degraded/uncertain
- limit distribution/amplification where harm is likely

Correction

- publish corrected explanations with diffs
- link corrections to original instances
- notify affected participants

Retrospective

- root cause (policy, model, UI, incentives)
- prevention changes (tests, thresholds, UI fixes)

Learning loops

- update conformance tests (Section 15)
- adjust thresholds and disclosure tiers

---

15. Path Toward ML-RFC (Conformance & Testing)

DP14 must be testable.

15.1 Conformance suites

- **Fidelity tests:** explanation vs. decision pathway
- **Consistency tests:** similar inputs → similar explanations
- **Deception tests:** selective disclosure, UI framing, AI narrative traps
- **Propagation tests:** export/import with degradation signaling

15.2 Reference implementations

- overlay panels with summary + drill-down evidence
- standardized explanation cards with confidence + provenance links
- dispute/appeal widgets bound to items

15.3 Data and artifacts

- explanation schemas (fields, types, confidence)
- provenance links (DP15) required for high-impact claims
- change logs for explanations (versioned)

15.4 Governance procedures

- SLA for dispute response and correction
- thresholds for mandatory external verification (high risk)
- zone-based escalation paths (DP8)

15.5 Promotion criteria

- ≥ target fidelity score across scenarios
- measurable reduction in misleading incidents
- verified cross-system preservation with explicit degradation
- functioning dispute → correction → retrospective loop

---

16. Closing Orientation (Operational Standard)

DP14 sets the **operational standard for belief formation** on the Meta-Layer.

A system is aligned only if a reasonable participant can:
- determine **what is known vs. uncertain**
- see **why a decision happened** and **verify it**
- understand **what incentives are in play**
- detect when signals are **degraded or contested**
- **challenge** and receive a **traceable correction**

Anti-goal:
- interfaces that are technically accurate but **systematically misleading** in practice

Standard:

> Signals must be **truthful, sufficiently complete, evidence-bound, and contestable**—and must remain so under pressure, incentives, and cross-system movement.

Trust is achieved when independent parties can **reproduce understanding** from the same signals and evidence.
