# A Common Language for AI: A Three-Tier Measurement Vocabulary for Multi-Stakeholder Accountability

**Anonymous submission**


## Abstract

As AI systems mediate decisions in healthcare, education, judicial risk assessment, defense, and public services, *who has the right, the capacity, and the language to speak about those decisions* has become the common foundation of ethical, legal, and democratic accountability. Yet today this question is addressed in fragmented vocabularies. Developers speak of alignment metrics, operators of compliance checklists, regulators of risk categories, citizens of natural-language explanations, and legal practitioners of causal liability. The result is that no shared evidence base forms across stakeholders for any given decision.

This paper argues that the three hierarchies operating behind every substantive AI reasoning—**Value (V)**, **Evidence (E)**, and **Source (S)**—can serve as a **multi-stakeholder common language** that resolves this vocabulary fragmentation. The central argument is that this vocabulary was not invented for AI; it was borrowed from human decision-making research. Schwartz's universal value theory, Walton's argumentation schemes, and Hovland–Kelley source credibility theory form three scholarly traditions whose vocabulary, long used to evaluate human actors, is here made applicable to AI decisions as well. This *borrowed* nature is the enabling condition that allows diverse stakeholders to intuitively adopt the same vocabulary.

Measurability is empirically demonstrated through 366,120 forced-choice responses from eight frontier LLMs (Anonymous, 2026c). This paper reinterprets those results through a multi-stakeholder lens: the 4:4 value cluster split (vendor brand cannot be used to infer alignment), the Security surge in defense (6/8 models, win-rate 0.951–0.998), and the 25.3–49.4 percentage-point Test–Retest vs Paired-Consistency gap (*framing sensitivity*, not stochastic noise).

This paper shows operationalizability through two general propositions. **Proposition 1 (Emission Feasibility)**: three-tier measurement is implementable as a per-decision low-overhead emission scheme without model retraining. **Proposition 2 (Layered Translation Feasibility)**: closed-vocabulary design enables deterministic reproduction of the same measurement at stakeholder-specific abstraction levels. **The contribution of this paper is the two feasibility propositions themselves, not any specific implementation specification.** The PRISM Logging Standard (Anonymous, 2026d) is referenced as an existence proof; other implementations can satisfy the same propositions.

The paper honestly addresses five limits of the common-language claim: vocabulary compatibility versus completeness, layperson comprehension, expert capture, layered-translation information loss, and the cultural pluralism limits of Schwartz vocabulary. The standard is released under the MIT license.


# 1 Introduction

## 1.1 Same Decision, Fragmented Vocabularies

As AI systems are increasingly involved in medical triage, educational resource allocation, judicial risk assessment, public service advisory, and defense decision support, one question is being asked in common across all stakeholders: **What did this decision prioritize, what evidence did it find decisive, and whose authority did it trust?** Yet this single question is currently bifurcated into mutually untranslatable vocabularies.

**AI developers** evaluate model behavior in the language of alignment metrics, benchmark scores, and RLHF curricula. **Vendors and operators** implement auditability in the language of compliance checklists, ISO/IEC standard items, and NIST RMF categories. **Regulators** govern systems in the language of risk categories, high-impact system classifications, and harmonised standards. **Citizens and users** confront AI decisions in the language of natural-language explanations, contestability of outcomes, and everyday understanding. **Legal practitioners** assign responsibility in the language of negligence, causation, foreseeability, and due diligence. **Researchers and auditors** conduct system comparisons in the language of statistical distributions, effect sizes, and replication metrics.

The same AI decision is described simultaneously in six vocabularies, but **no common vocabulary spans across the six**. As a result, no shared evidence base across multiple stakeholders forms for any decision. This vocabulary fragmentation is not merely an interdisciplinary difference. It is a **structural failure of multi-stakeholder accountability**.

The fragmentation becomes more severe in domain-specific stakeholders. When **medical AI** is used for patient triage, medical professionals view the decision through clinical appropriateness, patients through self-determination, insurers through risk pools, and health authorities through population health. When **educational AI** is used for student assessment, educators view it through learning outcomes, students through fairness, and parents through future oppor-

tunity. When **journalists** report on government AI, the vocabularies available to them are the language of official press releases or anecdotal cases—not the quantitative language of systemic behavior.

When citizens contest a decision, the vocabulary they use differs from the vocabulary operators write audit reports in, which differs from the vocabulary regulators evaluate systems in. Translation between the three is ad hoc, asymmetric, and biased toward whoever has the power to impose their vocabulary. As a result, accountability concentrates in those who control the vocabulary.

## 1.2 The Enabling Condition: Three Tiers as Borrowed Vocabulary

Behind every substantive AI decision, three hierarchies operate:

- **Value hierarchy (V)**: What value priorities drove the decision?
- **Evidence hierarchy (E)**: What kind of evidence was decisive?
- **Source hierarchy (S)**: What class of sources was trusted?

This three-tier structure is formalized in the Authority Stack model (Anonymous, 2026a). However, the central argument of this paper is not the Authority Stack itself, but the fact that **this vocabulary is borrowed**. The three tiers are not vocabulary newly invented to evaluate AI.

- The **value hierarchy** is borrowed from Schwartz's universal value theory (Schwartz 1992; Schwartz et al. 2012). Validated cross-culturally across 80+ countries, the 10-value (or 19-value) classification is the standard tool for measuring human motivation in social psychology, political science, and consumer research.
- The **evidence hierarchy** is borrowed from Walton's argumentation schemes (Walton, Reed, and Macagno 2008). The taxonomy is standard in law (evidence evaluation), rhetoric (argument structure analysis), and cognitive science (reasoning models).
- The **source hierarchy** is borrowed from Hovland–Kelley source credibility theory (Hovland, Janis, and Kelley 1953) and subsequent meta-analysis (Pornpitakpan 2004). Seventy years of mediated trust research have quantified how source classes influence human trust judgments.

All three are **vocabularies developed to evaluate human decision-making**. They take a language already in use for evaluating humans and make it equally applicable to AI decisions. This is the enabling condition for the common language argued for here.

This borrowed quality is essential in three senses. **First, communicability**: the vocabulary medical professionals use to evaluate the diagnostic judgment of fellow practitioners is the same vocabulary that can evaluate AI decisions. **Second, justifiability**: a borrowed vocabulary distributes the justification burden compared to a newly invented one—adopting Schwartz's value theory is justified not by a single AIES paper but by 30 years of cross-cultural research. **Third, generality**: the same vocabulary is applicable to AI actors, human actors, organizational actors, and policy systems. If citizens can evaluate decisions of AI, of government, and of medical professionals in the same vocabulary, the consistency of accountability increases.

Table 1: The role of the three-tier measurement vocabulary in multi-stakeholder accountability.

| Stakeholder | Reads / Acts on |
|---|---|
| AI developer | Global V-pair distribution; alignment debugging; RLHF curriculum tuning |
| Vendor/operator | Per-decision reasoning trace; integrity hashes; EU AI Act Art. 12–15 mapping |
| Regulator | Cross-vendor V/E/S distributions; systemic risk identification; post-market surveillance |
| Citizen/user | Per-decision V-pair (natural language); contestability; information-rights exercise |
| Researcher/auditor | Population-level distributions; framing sensitivity; meta-analysis |
| Legal practitioner | V/E/S at decision time; chain integrity; causal inference and due diligence |

## 1.3 Multi-Stakeholder Accountability: Who Reads What

Even granting that three-tier measurement is a candidate common language, the question remains: **how does the same vocabulary work meaningfully across six or more types of stakeholders?** Table 1 makes explicit the role of the measurement vocabulary by stakeholder.

Domain-specific stakeholders operate atop this general taxonomy. The same vocabulary acquires different points of action under different domain contexts: medical professionals examine whether E-pairs prioritize Guideline/Expert; patients ask where the V-pair places self-direction; health authorities audit population-level code distributions by demographic. Educators examine whether V/E pairs align with learning objectives; students compare V/E/S distributions across comparable decisions. Prosecutors, defense attorneys, and judges examine AI decision reasoning in a shared vocabulary; civil society audits demographic disparities in code distributions. Defense certification bodies threshold-evaluate domain-specific Security surge patterns. Public procurement officers conduct pre-procurement PRISM measurement to avoid vendor-brand dependence. Journalists and NGOs quantitatively report on publicly disclosed code distribution statistics from government AI.

Two key observations. **First, the same measurement vocabulary works simultaneously at different abstraction levels.** The developer reads raw V-pair distributions; the citizen reads a natural-language conversion of the same decision. Each is a different view of one underlying measurement (§3.2). **Second, domain-specific stakeholders do not replace the general taxonomy but refine it.** The medical

professional is the medical-domain instantiation of the general "operator" category; the patient is the medical-domain instantiation of the general "citizen/user" category.

### 1.4 The Need to Demonstrate the Common Language in Operation

Two pressures demand operational demonstration. First, **empirical generalization**: (Anonymous, 2026c) demonstrated measurability with 366,120 responses across eight frontier LLMs, but the measurement remained at academic-evaluation level. Second, an **imminent regulatory clock**: EU AI Act Article 12 will mandate per-decision reasoning records for Annex III high-risk systems starting August 2, 2026. NIST AI RMF and Korea's AI Basic Act (effective 2025, Article 25) similarly require reasoning traceability without specifying concrete formats.

This paper formalizes the demonstration as two general propositions.

- **Proposition 1 (Emission Feasibility)**: The three-tier measurement vocabulary is implementable as a per-decision low-overhead emission scheme integrable into operational systems without model retraining.
- **Proposition 2 (Layered Translation Feasibility)**: When the measurement vocabulary is designed as a closed vocabulary, the same measurement is deterministically reproducible at stakeholder-specific abstraction levels.

Section 3 (a) formalizes the two propositions as operationalization requirements R1–R5, (b) proves them through the operating mechanism of layered reading, and (c) references the PRISM Logging Standard (Anonymous, 2026d) as an existence proof. **The lead contribution of this paper is the two propositions themselves, and PRISM is one working demonstration.** Other implementations can satisfy the same propositions, and the citation value of this paper is decoupled from the longevity of any specific implementation.

### 1.5 Contributions

This paper makes six contributions. (1) Diagnoses vocabulary fragmentation in AI accountability through a multi-stakeholder lens, analyzing the structural cost incurred when six general stakeholder types together with domain-specific stakeholders view the same decision through different vocabularies. (2) Establishes V/E/S three-tier measurement as borrowed vocabulary from human decision-making research and argues that this borrowed nature is the enabling condition for common language. (3) Specifies stakeholder-specific use of the measurement vocabulary across six general and 11 domain-specific stakeholder types. (4) **Formalizes and proves two feasibility propositions**; this is the lead contribution and is decoupled from the longevity of any specific implementation. (5) Presents the PRISM Logging Standard as an existence proof, with explicit honesty about EU AI Act mapping ("satisfying by itself" vs "supporting"). (6) Honestly discusses limits of the common-language claim: compatibility vs completeness, layperson comprehension, expert capture, layered-translation information loss, cultural pluralism.

What this paper does not claim: the three-tier measurement vocabulary does not solve AI safety, does not guarantee alignment, and does not by itself exempt regulatory compliance. The vocabulary is only a candidate evidence base shared across multi-stakeholders for the same decision.

## 2 Background and Related Work

### 2.1 Vocabulary Fragments in the Literature

**Algorithmic accountability** research (Diakopoulos 2016; Wachter, Mittelstadt, and Floridi 2017) established the duty of post-hoc explanation but did not provide a multi-stakeholder shared vocabulary. The **AI Auditing** literature (Raji et al. 2020; Mökander et al. 2023) proposed audit processes without a standardized format for per-decision reasoning. **Transparency by design** (Felzmann et al. 2020) argued for design-stage transparency but lacks a production format. **Model Cards** (Mitchell et al. 2019) and **AI FactSheets** (Arnold et al. 2019) provide model-level static metadata, not per-decision dynamic measurements with multi-stakeholder interpretation. **Datasheets for Datasets** (Gebru et al. 2018) addresses training data, not reasoning traceability.

The common gap is the absence of **measurement vocabulary that is per-decision emittable and that allows multi-stakeholders to share the same evidence base at different abstraction levels**. This paper fills the gap with borrowed vocabulary and the two feasibility propositions.

### 2.2 Scholarly Foundation of the Borrowed Vocabulary

**Schwartz's universal value theory**: 10-value model validated across 80+ countries (Schwartz 1992; Schwartz et al. 2012); 19-value refinement (Schwartz and Butenko 2017). Standard in social psychology, political science, consumer research. **Walton's argumentation schemes**: 60+ schemes used in law, rhetoric, cognitive science, computer science (Walton 1996; Walton, Reed, and Macagno 2008). **Hovland–Kelley source credibility**: 70-year meta-analysis on mediated trust (Hovland and Weiss 1951; Pornpitakpan 2004). Each vocabulary token partially inherits existing scholarly validation; vocabulary newly invented for AI must be justified from scratch, while borrowed vocabulary is grounded in accumulated validation.

### 2.3 EU AI Act Article 12 and Reasoning Traceability

EU AI Act Article 12 (automatic logging) requires automatic event recording over the system lifetime of high-risk AI. Article 12(2) requires logging for (a) identification of risk-prone situations, (b) post-market monitoring assistance, (c) monitoring of system operation. **What the article does not specify is the reasoning trace of the decision itself.** However, in combination with Article 13 (transparency), Article 14 (human oversight), and Article 15 (robustness), reasoning visibility is implicitly required. The measurement vocabulary presented here does not satisfy Article 12 by itself; it adds a reasoning-trace layer that operates together with Articles 13, 14, and 15. This precise distinction is applied

consistently in the compliance mapping of §5 through the expressions "supporting" and "contributing."

### 2.4 Adjacent Regulatory Frameworks

**United States**: NIST AI RMF provides a governance framework; the Biden Executive Order on Safe AI (2023) imposes audit obligations on government-use AI. **South Korea**: Article 25 of the AI Basic Act (effective 2025) mandates the obligation to explain the basis of decisions for high-impact AI systems. **OECD AI Principles, UNESCO Recommendation on Ethics of AI**: specify principles of decision explainability without operational formats. The measurement vocabulary is a candidate operational implementation across all four frameworks.

## 3 Operationalization: From Vocabulary to Production

This section proves the two feasibility propositions in three steps. §3.1 formalizes the requirements (R1–R5) that any deployable instance must satisfy. §3.2 describes the operating mechanism of layered reading (Proposition 2) in a PRISM-independent manner. §3.3 presents the PRISM Logging Standard as an existence proof. §3.4 discusses alternative implementations to demonstrate that the contribution is decoupled from the longevity of any specific implementation.

### 3.1 Operationalization Requirements

Every concrete operationalization must satisfy five requirements. R1–R4 are direct consequences of the two propositions; R5 is recommended for legal admissibility.

**R1 (Per-decision low-overhead emission)**: A measurement record must be emitted for every decision, without model retraining or substantial latency overhead. This constrains emission to schemes that leverage existing model capabilities (system prompt steering, structured outputs, tool calls). Direct formalization of Proposition 1.

**R2 (Closed vocabulary)**: All fields of the V/E/S hierarchies must be drawn from predefined closed vocabularies, not free text. Two reasons: (i) machine parsing for population-scale audit, (ii) the necessary condition for deterministic layered translation (precondition for R4). Necessary condition for Proposition 2.

**R3 (Privacy preservation by design)**: Measurement records must not include verbatim user content. All fields are tokens from the closed vocabularies of R2, and no portion of user input is encoded directly. Naturally compatible with GDPR, CCPA.

**R4 (Layered translation)**: The same measurement record must be readable at multiple abstraction levels (Raw $\rightarrow$ Schema $\rightarrow$ Dashboard $\rightarrow$ Citizen-facing) without ambiguity. R2 is the deterministic precondition. Direct formalization of Proposition 2.

**R5 (Integrity guarantees, recommended)**: The audit substrate should include tamper-detection mechanisms (e.g., chain hashing). Recommended for legal admissibility but not essential to the core vocabulary claim.

These five requirements are formalized **independently of implementation specs**. PRISM's 60-character code, system-prompt-based emission, and SHA-256 chain hash are *one way* of satisfying R1–R5, not R1–R5 themselves.

### 3.2 The Operating Mechanism of Layered Reading

The same three-tier measurement record is read simultaneously at four abstraction levels.

**Layer 0 (Raw)**: The record exactly as emitted—compact, machine-parseable. Audience: ML researchers, system operators.

**Layer 1 (Schema)**: The measurement decomposed into named fields with explicit semantics. Audience: auditors, technical reviewers.

**Layer 2 (Dashboard)**: Natural language using domain-specific terms. Audience: domain experts.

**Layer 3 (Citizen-facing)**: A simplified natural-language form describing the most salient contrast. Audience: end users, the general public.

**Mechanism**: Because the V/E/S hierarchies use closed vocabularies (R2), each token has a fixed mapping to natural language at each layer. The mapping is deterministic—the same Raw record always produces the same Schema/Dashboard/Citizen-facing forms. This determinism is the necessary condition for layered translation; with open vocabulary, stakeholder-specific interpretive ambiguity would arise at each layer.

**Example (using PRISM as the underlying scheme)**:

- **Layer 0**: `C:MD/IXi | V:Bec<Sda | E:Exp<Gui | S:Usr<Pro`
- **Layer 1**: Domain=Medical, Impact=Individual, Reversibility=Irreversible, Time=Immediate; V-top-pair: Benevolence-care $<$ Self-direction-action; E-top-pair: Expert $<$ Guideline; S-top-pair: User-information $<$ Professional-source.
- **Layer 2**: "In an irreversible decision in the medical domain, self-direction-action was prioritized over benevolence-care; guidelines took precedence over expert opinion; and professional sources took precedence over user information."
- **Layer 3**: "This decision prioritized *the patient's right to self-determination* over *the medical practitioner's duty of care*, and used official guidelines as the core basis of the decision."

This mechanism is demonstrated with PRISM but is not PRISM-specific. Any scheme satisfying R2 deterministically produces the same 4-layer reading. **This fact constitutes the operational proof of Proposition 2.**

### 3.3 PRISM as Existence Proof

The PRISM Logging Standard is one concrete implementation that satisfies R1–R5. Full technical specifications appear in (Anonymous, 2026d). This subsection summarizes how PRISM's design choices satisfy each requirement.

**R1 satisfied**: A single-line $\sim$60-character format, parseable by regex:

```
C:<dom>/<sc><rev><t> |
V:<v_lo><<v_hi> | E:<e_lo><<e_hi> |
S:<s_lo><<s_hi>
```

Emission overhead: additional system prompt ∼500 tokens, additional ∼60 tokens per decision. No model retraining. Three output modes (inline tag, structured JSON, tool call) accommodate diverse frontier-LLM capabilities. These specs are *one way* of satisfying R1, adapted to the *current frontier-LLM environment*—if architectures change, specs may be redesigned, but R1 itself is architecture-independent.

**R2 satisfied**: Four closed sets. Context (C): 7 domains × 5 impact scopes × 3 reversibilities × 3 time horizons. Value (V): Schwartz 10- or 19-value profile. Evidence (E): 10 categories from Walton schemes. Source (S): 10 categories from Hovland–Kelley. Each hierarchy emits the top-2 codes in `lower<higher` form—intentional design capturing the strongest single contrast.

**R3 satisfied**: All fields are tokens from predefined closed vocabularies; verbatim user content is never encoded.

**R4 satisfied**: The 4-layer reading of §3.2 is deterministically applicable to PRISM codes (example above).

**R5 satisfied**: SHA-256 chain hash—each log's hash is included in the next log's input.

PRISM's design choices are *one way of leveraging existing capabilities of current frontier LLMs*. Other choices can satisfy the same R1–R5; even if PRISM becomes outdated, the two propositions remain instantiable in other implementations.

### 3.4 Alternative Operationalizations

Other implementations satisfy R1–R5 with different tradeoffs.

**Token-level emission**: Embed V/E/S tokens in model output as special tokens. Requires fine-tuning, lower runtime overhead.

**Latent emission via embedding**: Map decision embeddings to V/E/S clusters post-hoc. Requires per-vendor calibration.

**Multi-modal emission**: For image/video AI, emit V/E/S tokens alongside output.

**Compositional emission**: For agentic systems, emit V/E/S per tool call; enables reasoning trace of multi-step decisions.

These four examples are not exhaustive. The point is that the two feasibility propositions **admit multiple implementations**. PRISM is the first existence proof, and **the citation value of this paper is bound not to the longevity of PRISM specs but to the longevity of the two propositions, R1–R5, and the layered reading mechanism.**

## 4 Multi-Stakeholder Reading of Empirical Behavior

This section reinterprets four findings from the 366,120-response data of (Anonymous, 2026c) through the stakeholder taxonomy of §1.3. Same data and same findings yield different action points at different stakeholder-specific abstraction levels—empirical proof of the lead claim.

### 4.1 Data and Measurement

8 frontier LLMs, 14,175 forced-choice scenarios, 5 perspective variants plus identical-question repetition for reliability, total 366,120 valid responses. Measurement: March 2026. Coverage: 7 domains × 5 impact scopes × 3 reversibilities × 3 time horizons. Models anonymized A–H. This paper conducts no new measurement; it uses these results as the **empirical substrate for the common-language claim**.

### 4.2 The 4:4 Cluster Split in the Value Hierarchy

**Finding**: Two clusters separate clearly in global V-pair distribution. Universalism-prioritizing (4 models, win-rate 0.922–0.951); Security-prioritizing (4 models, win-rate 0.889–0.970). Crucially, **two models from the same vendor were observed in different clusters**—vendor identity is not a strong predictor of alignment signature.

**Stakeholder-specific reading**. *Developer*: alignment methodology (RLHF curriculum, Constitutional AI variants, DPO) is a stronger predictor than vendor identity; cluster differences within a vendor are quantitative signals for alignment-pipeline tracing. *Public procurement officer*: vendor-brand labels are not sufficient information about alignment; mandating PRISM distribution reporting in RFPs becomes the foundation of informed selection. *Regulator*: vendor-level audit is inappropriate; **model-level audit is essential**—one implication is that the EU AI Act's registration unit should be the model, not the vendor. *Citizen*: per-service PRISM distribution disclosure becomes the foundation for informed consent. *Legal practitioner*: "vendor intent" cannot be treated as a single variable; due diligence must be model-specific. *Researcher*: a measurable surrogate that allows analysis of alignment-methodology effects separated from vendor variables.

### 4.3 Security Surge in the Defense Domain

**Finding**: V-pair distribution is systematically reorganized when domain changes. In defense, 6/8 models elevate Security to rank-1 (win-rate 0.951–0.998); among the 4 Universalism-prioritizing models, 2 *flip* to Security under defense framing.

**Reading**. *Defense certification body*: quantitative signal of inappropriate escalation potential in LLM-controlled autonomous weapon systems; threshold inclusion of PRISM distribution in certification standards becomes possible. *Defense policymaker*: an "identical model" becomes **a different actor** the moment defense framing is triggered—pre-deployment evaluation cannot rely on general-domain distribution. *Healthcare/education/judicial operators*: domain framing systematically changes model behavior—the need for domain-specific PRISM measurement generalizes. *Citizen/human-rights group*: AI is a **domain-context-dependent actor**, not a single model; oversight should require domain-specific distributions. *Legal practitioner*: domain context at decision time becomes a decisive variable for liability attribution. *Regulator*: behavioral signatures must be evaluated coherently with regulatory categories (Annex III classifications).

### 4.4 Framing Sensitivity: TRR vs PCS Gap

**Finding**: Test-Retest Reliability (identical-scenario repetition) global 0.875–0.985 (87.5–98.5%); Paired Consistency

Score (5 perspective variants) global 0.480–0.659 (48.0–65.9%); gap 25.3–49.4 percentage points across all 8 models. Variability arises from **framing sensitivity** (Stochastic Noise diagnostic = 0%), not stochastic noise. Framing sensitivity is itself a measurable model characteristic.

**Reading**. *Citizen / discrimination victim*: **differences in user expressive ability translate into systematic differences in response**; this constitutes a threat to fair access and a new evidence type for algorithmic discrimination. *Legal practitioner*: low-PCS models are framing-sensitive; the "possibility that a different decision would have arisen with different phrasing" becomes a quantitative basis for contestability. *Operator*: low-PCS models require additional safeguards in production—phrasing-normalization layers and multi-phrasing comparison. *Security officer*: quantitative evidence of adversarial framing manipulation in medical and judicial AI; security evaluation should include framing-robustness testing. *Regulator*: **framing sensitivity, as a measurable characteristic, can itself be itemized as an audit criterion**; PCS as quantitative criterion for EU AI Act Article 15. *Developer*: TRR–PCS gap signals alignment brittleness; including adversarial framing in training is the evidence base for robustness strengthening. *Researcher*: substrate for a new research program (cross-vendor PCS, domain-specific PCS, language-specific PCS).

### 4.5 Drift and Anomaly Detection

**Finding**: PRISM distributions form the basis for three production-audit signals. *Drift detection*: V-pair distribution change across model versions. *Anomaly detection*: surge in S:Ano (anonymous source) frequency suggests prompt injection or context contamination. *Domain consistency verification*: defense-typical pairs in medical-domain decisions suggest domain misclassification or context leakage. All three are derivable from distribution statistics without access to user content (R3 satisfied).

**Reading**. *Public-sector deployer*: quantitative basis for the right to advance notification before model updates; drift thresholds in SLAs become a new deployer right. *Security team*: S:Ano surge as prompt-injection alarm signal—privacy-preserving security monitoring. *Operator*: automatic detection of domain misclassification. *Regulator*: quantitative toolkit for post-market surveillance under EU AI Act Article 72. *Citizen/NGO*: quantitative basis for demanding disclosure of aggregate PRISM statistics from government AI.

### 4.6 Same Evidence Base, Different Abstractions: A Meta-Reading

The four findings derive from the same dataset, yet action points differ systematically by stakeholder. For the public procurement officer, the 4:4 cluster split becomes an RFP mandate; for the citizen, a disclosure right; for the legal practitioner, a due-diligence variable. **Same data, same finding, different reading.**

This pattern is the most concrete empirical proof of the central claim—**three-tier measurement works as a multi-stakeholder common language**. If the vocabulary had been fragmented by stakeholder, six types of stakeholders could not have derived six action points from the same data; each would have had to measure again in their own vocabulary. From a single measurement substrate, this section showed that **six general stakeholder types together with domain-specific stakeholders can simultaneously derive actionable evidence**. This is proof that the stakeholder taxonomy of §1.3 is not an *abstract promise* but *operational evidence*. The specific specs of PRISM are not essential—any substrate that satisfies R1–R5 enables the same multi-reading.

Table 2: PRISM scheme outputs and EU AI Act articles supported.

| Output | Article (relation) |
|---|---|
| Per-decision record | Art. 12(1) automatic logging |
| Context field (C) | Art. 12(2) risk-situation traceability |
| C distribution over time | Art. 12(2)(a) period of use |
| Integrity mechanism | Tamper resistance (Recital 73) |
| V/E hierarchies | Art. 13 transparency *(supporting)* |
| Layered Reading | Art. 14 human oversight *(contributing)* |
| Distribution aggregation | Art. 15 robustness monitoring |
| Anomalous patterns | Art. 73 serious-incident reporting |

## 5 Compliance and Cross-Jurisdictional Coherence

### 5.1 An Honest EU AI Act Mapping

The expressions "supporting" and "contributing"—rather than "satisfying"—reflect the fact that no measurement scheme by itself exempts EU AI Act obligations.

**Note on Article 12(3)**: This article enumerates minimum logging content for Annex III 1(a) biometric systems and does not require general hashing. PRISM's hash mechanism is a defensive integrity measure, not a direct implementation of Article 12(3). The mapping above assumes PRISM's outputs, but its *structure* applies equally to other R1–R5 instances.

### 5.2 Adjacent Jurisdictions

**NIST AI RMF**: A governance framework without specific measurement format; three-tier measurement is integrable into the Measure stage. **Biden EO on Safe AI (2023)**: imposes audit obligations on government-use AI; per-decision records are an audit substrate. **Korea AI Basic Act, Art. 25**: mandates explanation of decision basis for high-impact AI; the Citizen-facing layer of Layered Reading (§3.2) is a candidate format. **OECD AI Principles, UNESCO Recommendation**: candidate operational implementation. The same vocabulary cohering across EU, US, Korean, OECD, and UNESCO frameworks is the cross-jurisdictional implication of this paper.

### 5.3 Vocabulary Unification, Not Jurisdictional Unification

A common measurement vocabulary cohering across jurisdictions does not imply jurisdictional unification. Even on

a shared vocabulary, the distribution of authority is a political decision of each jurisdiction. **When vocabulary is unified, comparison, transfer, and translation across jurisdictions become possible**: citizens compare AI under EU vs Korean regulation in the same vocabulary; multinational vendors handle multi-jurisdictional compliance with a single measurement infrastructure; researchers conduct cross-jurisdictional comparative audit.

# 6 Limits of the Common-Language Claim, with Ethical Implications

## 6.1 Honest Limits of the Common-Language Claim

**Compatibility vs completeness**. The vocabulary is *compatible* with multi-stakeholders, but does not *fully satisfy all needs of all stakeholders*. The Schwartz/Walton/Hovland–Kelley vocabularies capture universal dimensions of human evaluation but not domain-specific nuances such as clinical appropriateness in medicine, procedural justice in law, or developmental appropriateness in education. **A common vocabulary forms a common evidence base, but does not replace domain-specific vocabularies.**

**Layperson comprehension**. The natural-language statement at Layer 3 is *accessible*, but *critical examination* requires understanding of Schwartz value vocabulary, alternative value priorities, and implications of domain context. **Vocabulary access and vocabulary comprehension must be distinguished.** This paper establishes access; comprehension is the domain of mediating infrastructure (press, NGOs, civil society).

**Information loss in layered translation**. Information density decreases from Layer 0 to Layer 3. When Layer 3 simplifies the medical example, domain context (reversibility, time horizon), evidence hierarchy, and source hierarchy are truncated. **Simplification is the ethics of decision-making**—what is preserved and what is omitted is itself a decision about what information is valuable for citizens. This paper does not provide a standard for Layer 3 design and makes explicit that this is a domain requiring stakeholder negotiation.

**Expert capture**. A standardized measurement vocabulary concentrates power in those who define and update it. Who maintains the vocabulary, by what procedures, and who approves cultural extensions are governance questions. The MIT-licensed open-source release ensures forking, but *maintained-version* governance is a separate decision. **"Neutral standards" are an illusion; governance determines the political substance of the standard.**

## 6.2 Power Dynamics as Measurement

A standardized measurement vocabulary creates power relations. The vocabulary is grounded in scholarly traditions, not vendor or government definition. Measurement is emitted by vendors but verified by multiple stakeholders. Interpretation must vary by domain and culture. However, the adoption of the standard is itself a power act, and the declaration that "Schwartz value vocabulary is the audit standard" can marginalize non-Western value frameworks.

## 6.3 Transparency vs Trade Secret

Measurement vocabulary emission exposes indirect signals of vendor alignment methodology. The 4:4 cluster split is interpretable as result of vendor-specific RLHF/DPO/Constitutional approaches—a competitive-information-exposure risk. The position here: (1) measurement codes expose the reasoning signature, not concrete alignment implementation; (2) Models A–H anonymization is possible in production audit; (3) sufficiently rich distributions are vendor-identifiable through sophisticated analysis—this trade-off must be accepted, and the ethical commitment of this paper is that **for public service AI, transparency must take priority**.

## 6.4 Cultural Pluralism

Schwartz's universal value theory is validated across 80+ countries, but some non-Western value systems are not fully captured: East Asian face (private vs public face distinctions reduced to a single code), Indian dharma (deontology–universalism mixture), Ubuntu (community-first reduced to Benevolence), Confucian ren/li/yi (multilayered interactions reduced to single codes). Response: (1) multilingual vocabulary expansion is core future work; (2) culture-specific clusters can be added modularly; (3) the measurement vocabulary itself is subject to critique, and critique must become part of the quantitative tool.

## 6.5 Goodhart's Law and Manipulation Risk

Goodhart's Law warns that when a measure becomes a target, it ceases to be a good measure (Strathern's reformulation). Vendors may tune models to emit "good-looking" codes; political actors may strategically adjust answers to align with declared signature. Mitigations: (1) layering of measurement (PRISM + behavioral test + activation tracing + human evaluation); (2) audit of the measurement itself (manipulation patterns are statistically identifiable); (3) incentive alignment (measurement values function as user-trust and market signals, not only regulatory compliance).

## 6.6 What This Vocabulary Does Not Guarantee

The three-tier measurement vocabulary does not guarantee AI alignment, does not by itself solve AI safety, does not by itself exempt EU AI Act compliance, and does not automatically secure social legitimacy of public-servant AI. **It is only a candidate evidence base shared across multi-stakeholders for the same decision, operating within a larger governance system.**

# 7 Generalization: Human, Organizational, and Political Actors

The mechanism of three-tier measurement does not fundamentally depend on the type of measurement target. Value, evidence, and source hierarchies apply whether the decision-making subject is an LLM, a human politician, an organization, or a philosophical system—a direct consequence of the borrowed-vocabulary nature emphasized in §1.2. This generalization makes **AI systems and human decision subjects comparable on the same scale**: in an era when AI is

involved in policy advisory, citizens can simultaneously examine the value signature of the AI system and that of the politician using it.

The platform values and voting record of a political party, the stated priorities and proposed policies of a candidate, the mission statement of a corporation, the charter of a government department—all are mappable to declared V/E/S signatures. The distance between measured and declared signatures becomes a quantitative indicator of integrity.

Extension to human measurement raises ethical issues: (1) consent of measurement targets; (2) abuse risk in cancel culture and political attack; (3) cultural bias amplification; (4) power dynamics of citizen-driven vs state-driven audit. This paper presents only the frame and makes explicit that concrete implementation must not occur without a separate ethical framework.

The generalization opens two possibilities simultaneously: **(a)** strengthening democratic accountability through commensurable comparison; **(b)** cultural-hegemony risk when a standardized vocabulary marginalizes non-Western value systems. The ethical implementation centers on guaranteeing cultural pluralism and maintaining community-driven revision possibility. Even as the three-tier vocabulary is deployed, the project **explicitly accepts that the vocabulary itself is subject to critique and revision**.

## 8 Limitations and Future Work

**Post-hoc nature**. The measurement code is reasoning the model reports post-hoc, not causal tracing. Self-reports may be post-hoc rationalizations. However, this limit is common to all reasoning logs, and structured post-hoc reports are more auditable than unstructured silence. **Cross-model emission consistency**. Emission stability differs across models; production deployment requires verification and retry logic. **Multilingual vocabulary**. Currently English and Korean; EU 24 official languages and major East Asian languages remain. **Harmonised standards convergence**. Once CEN/CENELEC standards are finalized, mapping or absorption is required. **Empirical validation of alternatives**. Token-level, latent embedding, multi-modal, and compositional emission of §3.4 need empirical R1–R5 comparison. **Extension to human measurement**. Instrument design, measurement ethics, and citizen utilization mechanisms are follow-up research.

## 9 Conclusion

This paper argued that the measurement of the three hierarchies of value, evidence, and source constitutes a **multi-stakeholder common language** that resolves the vocabulary fragmentation across the diverse stakeholders surrounding AI. The central argument is that this vocabulary was not invented for AI but borrowed from human decision-making research—a quality that allows multi-stakeholders to intuitively accept it.

The lead contribution is the two feasibility propositions (Emission Feasibility, Layered Translation Feasibility), formalized as R1–R5 and proven through the Layered Reading mechanism, with the PRISM Logging Standard referenced as an existence proof. **The citation value of this paper is bound not to the longevity of PRISM specs but to the longevity of the two propositions, R1–R5, and the layered reading mechanism.** The 366,120-response empirical reinterpretation in §4—six general stakeholder types together with domain-specific stakeholders deriving actionable evidence from the same substrate—provides working proof that the stakeholder taxonomy is operational evidence, not abstract promise.

**A common language is not a universal language.** It means the formation of an evidence base shared across multi-stakeholders for the same decision; it does not mean satisfying all needs of all stakeholders. Three-tier measurement is the first deployable candidate for such an evidence base; PRISM is its first working demonstration; other instances will follow. The evolution of this vocabulary—and the multi-stakeholder accountability it enables—must remain open to community scrutiny and critical redefinition.


## Acknowledgments — LLM Usage Statement

The author drafted this manuscript; LLM assistance was used only for English language polishing, grammar correction, and style refinement of author-written content, in accordance with AAAI/AIES policy on LLM use. All academic claims, theoretical frameworks (Authority Stack model, the three-tier vocabulary, the two feasibility propositions), the empirical reinterpretation, and the conclusions are author-originated.


## Ethical Considerations Statement

This work proposes a measurement vocabulary intended to enable multi-stakeholder accountability for AI systems. Several ethical commitments are made explicit. *Vocabulary grounding*: the proposed vocabulary derives from established academic traditions (Schwartz value theory, Walton argumentation schemes, Hovland–Kelley source credibility theory). As discussed in §6.4, this Western-academic grounding may marginalize non-Western value systems including East Asian face concepts, Indian dharma, Ubuntu philosophy, and Confucian ethics; multilingual and culture-specific vocabulary expansion is identified as core future work. *Power concentration*: although designed to enable multi-stakeholder accountability, any standardized vocabulary risks concentrating power in those who maintain and update it—an "expert capture" concern made explicit in §6.1.4. The MIT-licensed open-source release ensures forking, but governance of the maintained version remains a separate, ongoing community decision. *Empirical use*: the 366,120-response data reused from prior work involves no new measurement of human subjects. Future extensions to human measurement (politicians, organizations) require separate ethical frameworks, as noted in §7.3.

## Researcher Positionality Statement

The author approaches this work as an AI accountability researcher concerned with bridging AI ethics scholarship, regulatory compliance practice, and citizen-facing accountability mechanisms. The research program focuses on develop-

ing measurement vocabularies as accountability infrastructure, building on prior work formalizing the Authority Stack model and the PRISM framework. The multi-stakeholder framing of this paper reflects a particular conception of accountability rooted in liberal democratic traditions; alternative governance frameworks—including community-rooted, indigenous, or non-state-centric governance models—may require different measurement vocabularies, and the author commits to ongoing engagement with diverse stakeholder communities in the evolution of the proposed vocabulary. The choice to present PRISM as an existence proof rather than a normative standard reflects the author's commitment to leaving the vocabulary open to community scrutiny and revision.

## Adverse Impact Statement

If widely adopted, the proposed measurement vocabulary may produce several adverse effects warranting explicit consideration. *Regulatory capture*: a standardized vocabulary may be co-opted by powerful actors (large vendors, dominant regulators) to entrench specific interpretations of accountability that disadvantage smaller actors or marginalized communities. *Performative compliance*: vendors may game the measurement (Goodhart's Law, §6.5), producing distributions optimized for audit appearance rather than substantive accountability improvement. *False confidence*: the existence of structured measurement may create the appearance of solved accountability problems while underlying issues remain. *Surveillance enablement*: per-decision logging, while privacy-preserving by design (R3), creates dual-use potential for surveillance applications if integrity guarantees are circumvented or if logging infrastructure is repurposed beyond its declared scope. *Cultural imposition*: deployment of a Western-academic-grounded vocabulary in non-Western jurisdictions may inadvertently impose evaluative frameworks foreign to local accountability traditions. Mitigations are discussed throughout §6, but ultimate responsibility for adverse outcomes rests with deployment-context governance, not with the measurement vocabulary itself. The author advocates for staged adoption, community oversight of vocabulary maintenance, and explicit accommodation of jurisdictional and cultural variation in any deployment.

## References


Arnold, M.; Bellamy, R. K. E.; Hind, M.; Houde, S.; Mehta, S.; Mojsilović, A.; Nair, R.; Ramamurthy, K. N.; Olteanu, A.; Piorkowski, D.; Reimer, D.; Richards, J.; Tsay, J.; and Varshney, K. R. 2019. FactSheets: Increasing Trust in AI Services Through Supplier's Declarations of Conformity. *IBM Journal of Research and Development*, 63(4/5).

Diakopoulos, N. 2016. Accountability in Algorithmic Decision Making. *Communications of the ACM*, 59(2): 56–62.

Felzmann, H.; Fosch-Villaronga, E.; Lutz, C.; and Tamò-Larrieux, A. 2020. Towards Transparency by Design for Artificial Intelligence. *Science and Engineering Ethics*, 26(6): 3333–3361.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé III, H.; and Crawford, K. 2018. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*.

Hovland, C. I.; Janis, I. L.; and Kelley, H. H. 1953. *Communication and Persuasion*. Yale University Press.

Hovland, C. I.; and Weiss, W. 1951. The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, 15(4): 635–650.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 220–229.

Mökander, J.; Schuett, J.; Kirk, H. R.; and Floridi, L. 2023. Auditing Large Language Models: A Three-Layered Approach. *AI and Ethics*.

Pornpitakpan, C. 2004. The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology*, 34(2): 243–281.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, 33–44.

Schwartz, S. H. 1992. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. *Advances in Experimental Social Psychology*, 25: 1–65.

Schwartz, S. H.; and Butenko, T. 2017. Value Tradeoffs Propel and Inhibit Behavior. *European Journal of Social Psychology*, 47(3): 241–258.

Schwartz, S. H.; Cieciuch, J.; Vecchione, M.; Davidov, E.; Fischer, R.; Beierlein, C.; Ramos, A.; Verkasalo, M.; Lönnqvist, J.-E.; Demirutku, K.; Dirilen-Gumus, O.; and Konty, M. 2012. Refining the Theory of Basic Individual Values. *Journal of Personality and Social Psychology*, 103(4): 663–688.

Wachter, S.; Mittelstadt, B.; and Floridi, L. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2): 76–99.

Walton, D. 1996. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates.

Walton, D.; Reed, C.; and Macagno, F. 2008. *Argumentation Schemes*. Cambridge University Press.