Three-Layer Measurement as a Common Language for AI Integrity: The PRISM Logging Standard

Anonymous Authors[1]

Abstract

As AI systems mediate decisions in public institutions, examining the reasoning behind those decisions has become a shared challenge across the AI safety, AI for social good, and AI governance communities. Yet these communities use different vocabularies to address the same problem. **We argue that the three hierarchies operating behind every AI decision—value, evidence, and source—provide a unified measurement language for AI integrity at large.**

This vocabulary was formalized in the Authority Stack model (**?**) and empirically validated through 366,120 forced-choice responses across 8 frontier AI models (**?**). The same measurements suggest output prediction and anomaly detection are tractable (**?**). Three-layer measurement supports cross-model comparison, domain-specific behavioral analysis, and drift detection; we argue the same vocabulary generalizes to integrity evaluation of human individuals, organizations, political actors, and philosophical systems.

We introduce the PRISM Logging Standard as the first deployable demonstration of this generalization claim. Each substantive decision produces a structured ˜60-character code, emitted by major LLMs through a system-prompt addition alone. We provide three output modes (inline tag, structured JSON, tool call) and two value profiles (Schwartz 1992 ten-value, Schwartz 2012 nineteen-value). As a first test case, we present a compliance mapping to EU AI Act Article 12. The standard is released under MIT license.

Three-layer measurement is not a partial tool; it is the foundation of integrity evaluation infrastructure for the AI era. We support this argument through (i) establishing the measurement vocabulary, (ii) empirical demonstration, (iii) the first deployment instance, and (iv) a generalization path toward humans, organizations, and democratic society.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

1.1. The Three-Languages Problem

As AI systems mediate decisions in domains such as healthcare triage, education funding, and judicial risk assessment, examining the reasoning behind those decisions has become a central question of AI governance. Yet this question is currently addressed in three separate communities.

AI safety researchers focus on model-level alignment failures and risky behavior. AI-for-social-good researchers evaluate unintended harms and benefits at population scale. AI governance researchers design regulatory frameworks and accountability mechanisms. The three communities address the same fundamental question—how to make model behavior visible, comparable, and accountable—but use different vocabularies to do so.

This vocabulary fragmentation has practical costs. When a public institution deploys an AI system, the safety researcher's evaluation, the social-good evaluator's impact assessment, and the governance advisor's compliance memo do not converge on a shared evidence base. Citizens demanding accountability, engineers operating the system, and authorities regulating it see the same decision in different languages. The result is partial visibility from each angle, with no integrated accountability framework.

1.2. Common Foundation: Three-Layer Measurement and Its Empirical Basis

This fragmentation is not an essential limitation of AI; it stems from the absence of measurement vocabulary. Behind every substantive AI decision, three hierarchies operate:

- **Value hierarchy**: Which value priorities led the decision?

- **Evidence hierarchy**: Which kinds of evidence were decisive?

- **Source hierarchy**: Which source classes were trusted?

This three-layer structure was formalized in the Authority Stack model (**?**). It integrates measurable dimensions from three established research traditions: Schwartz's universal value theory (**??**), Walton's argumentation schemes (**?**), and Hovland-Kelley's source credibility theory (**??**). Every AI decision can be described in terms of these three hierarchies.

Our central claim follows: **the three hierarchies are measurable, and standardizing this measurement creates a common language that lets the AI safety, social-good, and governance communities operate from a shared evidence base.**

The claim is empirically supported. **?** measured the value, evidence, and source hierarchies of 8 frontier AI models. Measurement used forced-choice scenarios designed across 7 domains, 15 severity levels (combining 5 scope categories from individual to society with 3 reversibility classes), and 3 time horizons. To verify measurement reliability, the same scenarios were collected with 5 perspective variants (for Paired Consistency Score calculation) and exact repetitions (for Test-Retest Reliability calculation), yielding 366,120 total responses. Key findings:

- Models cluster into two distinct groups by value hierarchy—Universalism-first (4 models) and Security-first (4 models).

- Six of eight models inflate Security to 95.1–99.8% in defense domains—a domain-specific behavioral signature.

- Measurable cross-model differences also exist in evidence and source hierarchies.

- Reliability validation shows high test-retest reliability (TRR 91.7–98.6%) with substantially lower variant consistency (PCS 57.4–69.2%). The large gap between these measures indicates that variability arises from framing sensitivity rather than stochastic noise—and that this framing sensitivity is itself a measurable dimension of model behavior.

The three layers are measurable. This is not laboratory curiosity but a robust signal derived from production-relevant systems. However, the measurement currently lives in academic evaluation; it is not produced at deployment scale in real time.

1.3. Standardization and Our Contributions

Moving from academic evaluation to production audit requires two conditions: a compact format emittable per decision, and a vocabulary comparable across models and vendors. The need is accelerated by an imminent regulatory deadline. From August 2, 2026, EU AI Act Article 12 (**?**) mandates that operators of high-risk AI systems listed under Annex III maintain per-decision reasoning records. Yet official harmonised standards (CEN/CENELEC) remain under development, leaving operators to construct ad hoc tooling during the standards-gap window. These two pressures—generalizing academic measurement to production scale, and the imminent regulatory deadline—demand the same answer: **a standardized output format for three-layer measurement.**

Our contributions are fourfold:

1. **Three-layer measurement as a unified language.** We argue that the value, evidence, and source hierarchies operating behind every substantive AI decision are measurable, and that this measurement constitutes a unified evidence base capable of bridging the AI safety, social-good, and governance communities. We diagnose the absence of measurement vocabulary as the cause of community fragmentation, and identify three-layer measurement as the foundation of its resolution.

2. **Compressed empirical foundation.** Reframing the 366,120-response data from 8 frontier models (**?**) within our framework, we show that the three hierarchies (a) differ measurably across models, (b) form behavioral signatures under domain shifts, and (c) are usable for output prediction and anomaly detection (supplemented by **?**).

3. **First deployable instance—the PRISM Logging Standard.** We release a standardized output format that makes the measurement vocabulary emittable at production scale. Each substantive decision produces a ˜60-character structured code:

   ```
   C:MD/IXi | V:Bec<Sda
            | E:Exp<Gui | S:Usr<Pro
   ```

   The standard requires only a system-prompt addition; major LLMs emit it without retraining. We provide three output modes (inline tag, structured JSON, tool call) and two value profiles (Schwartz 1992 ten-value, Schwartz 2012 nineteen-value). As a first test case, we map the standard onto Articles 12, 13, 14, 15, and 73 of the EU AI Act, explicitly distinguishing what PRISM *supports* from what it does not single-handedly satisfy. The standard is released under MIT license.

4. **Generalization beyond AI.** We argue that the same three-layer measurement applies to integrity evaluation of human individuals, organizations, political actors, and philosophical systems (Section 7). Our claim is that three-layer measurement is not merely an AI compliance tool but **the foundation of integrity evaluation infrastructure for the AI era**. The PRISM Logging Standard is the first working demonstration of this foundation.

We are also explicit about what this paper does *not* claim. PRISM does not solve AI safety, does not guarantee alignment, and does not single-handedly discharge regulatory compliance. The standard is a deployment instance of the measurement vocabulary and operates within a larger governance system. Yet *the measurement vocabulary itself can be the common language for evaluating AI, humans, organizations, and societies*—this is our central claim.

2. Background

2.1. The Actual Requirements of EU AI Act Article 12

EU AI Act Article 12 (**?**) mandates automatic logging in high-risk AI systems. The article's explicit text comprises:

Article 12(1) requires automatic recording of events over the lifetime of the system. **Article 12(2)** requires logging capabilities enabling (a) identification of situations that may produce risk, (b) facilitation of post-market monitoring (Article 72), and (c) monitoring of system operation (Article 26(5)). **Article 12(3)** specifies minimum logging contents only for biometric identification systems under Annex III point 1(a), including period of operation, reference databases, matching inputs, and identification of verifying personnel.

What the article *explicitly* requires is event logging, traceability, and operational period. What it does *not* explicitly require is reasoning trace per se—"which values, evidence, and sources led the decision" appears nowhere in the Article 12 text. However, when combined with Article 13 (transparency), Article 14 (human oversight), and Article 15 (robustness), reasoning visibility is implicitly required. PRISM does not single-handedly satisfy Article 12's explicit requirements; it adds a reasoning-trace layer that operates alongside Articles 13/14/15. Integrity protection is implied by Recital 73 and the general traceability principle, not by Article 12(3) directly.

This precise distinction is consistently applied in the compliance mapping (Section 4) using "supports" and "contributes to" rather than "satisfies."

2.2. The Authority Stack and Measurement Vocabulary

The Authority Stack model (**?**) describes decision authority as hierarchized across four layers—output (L1), source (L2), evidence (L3), and value (L4). The upper three layers have measurable hierarchical structure, and measurement of these three enables partial prediction and verification of output behavior (**?**).

The measurement vocabulary integrates three research traditions. **Schwartz's universal value theory** (**??**) provides a 10-value (or refined 19-value) classification validated across 80+ countries. **Walton's argumentation schemes** (**?**) classify 10 evidence types used in reasoning. **Hovland-Kelley source credibility theory** (**?**), meta-analyzed in **?**, presents a credibility hierarchy for source classes. Each tradition has been independently validated; together they cover the measurable dimensions of AI decisions.

This paper converts that measurement vocabulary into a production-environment output format.

2.3. Related Work

PRISM does not replace the existing AI transparency and accountability tool ecosystem; it fills a gap between adjacent tools. Table 1 summarizes the differentiation.

Chain-of-Thought exposes reasoning as text but does not provide a structured audit format. Aggregating or comparing the CoT outputs of two models is difficult. PRISM captures post-hoc reported reasoning in a structured form—not causal trace, but the explicit trade-off enables comparability and aggregation. **Model Cards and FactSheets** are model-level static documents, complementary to per-decision dynamic logging. **NIST AI RMF** is a governance framework that does not prescribe a specific log format; PRISM is compatible as its execution layer. **OpenTelemetry** and **CloudTrail** provide traceability infrastructure as standard event logging; PRISM operates as the reasoning layer above this foundation.

The gap PRISM fills is clear: per-decision structured reasoning trace using a vocabulary comparable across models and vendors.

3. The PRISM Logging Standard

This section presents the specification of the PRISM Logging Standard, which converts three-layer measurement as a unified language into a production-environment output format. The design has two goals: (1) a compact format emittable per decision, and (2) a vocabulary comparable across models and vendors.

Table 1. PRISM compared with adjacent transparency and accountability tools.

Tool	Recording unit	Format	Cross-model comparability	Relation to PRISM
Model Cards (**?**)	Model	Static document	Limited (free-form)	Complementary (model-level)
AI FactSheets (**?**)	Model	Static document	Limited	Complementary (supplier conformity)
NIST AI RMF (**?**)	Governance	Framework	No format prescribed	Compatible (PRISM as execution)
Chain-of-Thought	Decision	Free text	Difficult (unstructured)	Complementary (causal trace)
OpenTelemetry / CloudTrail	Event	Key-value / structured	Infrastructure standard	Foundation (PRISM as reasoning)
PRISM Logging Standard	**Decision**	**Structured code**	**Direct cross-model**	—

3.1. Code Grammar

A PRISM code is produced as a single line per substantive decision:

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

The < symbol reads as "outranked by"—left is deprioritized, right is prevailed. The code is approximately 60 characters, single-line, and parseable via regular expression. Compactness is intentional, reducing storage cost and enabling SQL-friendly aggregation. Example:

```
C:MD/IXi | V:Bec<Sda
       | E:Exp<Gui | S:Usr<Pro
```

= Healthcare / Individual / Irreversible / immediate; Benevolence-Caring outranked by Self-Direction-Action; Expert opinion outranked by Authoritative Guideline; User-provided outranked by Professional source.

3.2. Vocabulary

The PRISM code uses four closed vocabulary sets.

Context (C): 7 domains (healthcare, education, legal, defense, finance, technology, general), 5 scope levels (individual to society), 3 reversibility levels (reversible, partial, irreversible), 3 time horizons (immediate, short-term, long-term). Time horizon is interpreted domain-relative—the "immediate" of healthcare and the "immediate" of legal are distinct time scales, with separate domain-specific clocks defined.

Value (V): A choice between two profiles. The **ten-value profile** (**?**) is robust in cross-model emission stability, while the **nineteen-value profile** (**?**) provides finer audit signals—enabling, for example, separation of self-direction into thought versus action, or power into dominance versus resources. Operators select one profile per system; logs from the two profiles are not directly comparable since their vocabulary sets differ.

Evidence (E) and **Source (S)**: 10 each. Evidence types are based on Walton's argumentation schemes (**?**) (from systematic review to anecdotal); source types are based on the Hovland-Kelley credibility hierarchy (from peer-reviewed academic to anonymous online).

For each hierarchy (V, E, S), Top 2 codes are emitted in `lower<higher` format—an intentional design that captures the strongest single opposition in a decision. Top 3+ loses cross-model consistency; Top 1 retains only single-value emphasis with diminished audit value.

The complete code definitions, cluster descriptions for vocabulary sets, and domain-specific time-scale tables are given in Appendix A.

3.3. Three Output Modes

Three modes are provided to accommodate model capabilities.

Mode A (Inline tag): Emitted within model output text, wrapped in `<prism_log>...</prism_log>` tags. The host application strips these tags before user exposure. A fallback for models that do not support structured output or tool calling.

Mode B (Structured JSON): A JSON object of form `{"response": "...", "prism_log": {"code": "..."}}`. Compatible with OpenAI Structured Outputs, Anthropic JSON, and Gemini JSON modes. Separation is guaranteed by the format itself.

Mode C (Tool call): The model invokes a dedicated `record_prism_log` tool. The response contains a separate tool-use block. Compatible with Anthropic Claude tool use, OpenAI function calling, and Gemini function calling. This is the cleanest separation and is recommended for agentic systems.

Critical UX rule: Across all modes, the PRISM log is never exposed to the end user. The log is routed only to audit storage.

3.4. Validation and Integrity

Syntactic validation: All PRISM codes can be checked for format integrity via a single regular expression. Vocabulary integrity is additionally guaranteed by validation against the

closed set of the chosen profile (10v or 19v).

Storage integrity: SHA-256 chained hashing—the hash of each log is included in the next log's input, so any tampering breaks the chain. This is a defensive integrity measure satisfying the general traceability principle (implied by EU AI Act Recital 73), not a direct implementation of Article 12(3).

Profile separation: 10v / 19v logs are regex-compatible but their vocabulary sets differ, so they are not directly comparable. Operators select one profile per system and maintain it.

The specific regex pattern, SHA-256 chaining algorithm, and validation code samples are given in Appendix B.

3.5. Privacy-Preserving Design

PRISM codes contain no verbatim user content. All fields are tokens from a pre-defined closed vocabulary; no part of user input is directly encoded. The C: domain metadata exposes only topic-level information (e.g., "this decision was a healthcare query"), which is at or below the level already exposed by existing routing logs. The V/E/S hierarchies are reasoning structure, not response content.

This characteristic enables the coexistence of population-scale audit and individual privacy. Distributional analysis over millions of decisions is possible without exposing any individual user's queries or responses to the log. The design naturally aligns with data protection regulations such as GDPR and CCPA, and is essential when population-scale behavioral audit serves as core social-good infrastructure—for example, evaluating citizen impact of public-service AI.

4. Compliance Mapping

This section maps the PRISM Logging Standard, as a unified measurement language, onto its first test case—the EU AI Act—with deliberate honesty: this is where general-language standardization meets concrete regulation.

Table 2 indicates which provisions PRISM *supports* or *contributes to*. The choice of "supports" and "contributes to" rather than "satisfies" is intentional weakening: PRISM logs do not single-handedly discharge any obligation.

Article 12(3) honesty: This subsection enumerates minimum logging contents specifically for biometric identification systems under Annex III 1(a) and does not require general hashing. Our hash tool is a defensive integrity measure, not a direct implementation of Article 12(3).

Articles 13/14 honesty: The primary frame of Article 13 is deployer-facing transparency; the core requirement of Article 14 is human stop/override capability. PRISM *supports* both articles' ancillary requirements but does not single-handedly fulfill their core requirements.

PRISM logs are a structured evidence layer that compliance programs submit to auditors and notified bodies alongside their own governance documentation, conventional event logs, and risk management documentation.

5. Empirical Behavior

5.1. Data: Authority Stack Mapping

This section reinterprets, in PRISM code form, the measurement data described in Section 2.1. The basis is 366,120 forced-choice responses from 8 frontier AI models across 7 domains, 15 severity levels (5 scope $\times$ 3 reversibility), and 3 time horizons (**?**). Measurement reliability was verified via 5 perspective variants (for PCS) and exact repetitions (for TRR), yielding TRR 91.7–98.6% and PCS 57.4–69.2%.

5.2. Conversion to PRISM Codes

The data of **?** was collected as ranking per layer (value, evidence, source). Conversion to PRISM codes is mechanical:

- Top 2 ranks per layer $\rightarrow$ V/E/S pair (`lower<higher`)
- Scenario domain $\rightarrow$ C: domain
- Scenario design (scope, reversibility, time) $\rightarrow$ rest of C: layer

Same data, different format. The conversion loses information only at Top 3 and below—an intentional trade-off (see Section 3.2).

5.3. Cross-Model Distribution Patterns

V-pair distributions divide cleanly into two clusters (**?**). Four models predominantly emit pairs in which Universalism outranks Security; the other four show the inverse pattern. This 4:4 split is interpreted as alignment-methodology variation (**?**). Measurable cross-model differences also exist in E- and S-pairs, with broad convergence on institutional-source-over-personal-source preference at S (**?**). In our framework, this means: while value signatures vary across models, institutional bias in source trust is shared.

V-pair distributions reorganize under domain shifts. The most striking case: in defense domains, 6 of 8 models inflate "any-value-outranked-by-Security" pair frequency to 95.1–99.8% (**?**). Domain-specific behavioral signatures form in other domains as well, as reported in the original dataset. PRISM code distributions are a natural format for expressing such signatures—the dataset of **?**, which holds the data as layer rankings, can be reinterpreted in our framework as code-pair distributions.

Table 2. PRISM elements and the EU AI Act provisions they support or contribute to.

PRISM element	EU AI Act provision (supports)	Auditor-facing usefulness
One PRISM line per decision	Art. 12(1) automatic record-keeping	Reasoning trace, complementing event logs
C: layer	Art. 12(2) traceability of risk situations	Per-decision risk-context tag
Aggregate C: distribution	Art. 12(2)(a) period of use (with timestamps)	Volume and category over time
SHA-256 chained hash	Tamper-evidence (implied by Recital 73)	Chain-of-custody evidence
V: and E: hierarchies	Art. 13 transparency *(supports)*	Reported value and evidence priorities
Single-line structured code	Art. 14 human oversight *(contributes)*	Auditable format at scale
Aggregate code distribution	Art. 15 accuracy and robustness monitoring	Drift signal across versions
Anomalous code patterns	Art. 73 serious incident reporting	Pre-investigation signal
S: source hierarchy	Art. 10 data governance *(supporting)*	Per-decision source class record

5.4. Drift and Anomaly Detection Signals

The dataset is a single-time-point measurement, but the following signals are immediately actionable when PRISM is deployed at production scale:

- **Drift detection**: Changes in V-pair distribution across model versions. A sudden shift from Universalism-first to Security-first signals an alignment change.
- **Anomaly detection**: Frequency of `S:Ano` (anonymous source) outranking other sources. A sudden surge raises suspicion of prompt injection or context contamination.
- **Domain consistency check**: Defense-typical pairs (e.g., any-value-outranked-by-Security) appearing in healthcare decisions indicates domain misidentification or context leakage.

All three signals are derivable from code-distribution statistics; they require neither separate model analysis nor access to user content.

5.5. Implications

The three hierarchies are stably emittable across diverse models and domains, and behavioral signatures under domain shifts are measurably distinct, as confirmed in the data of **?**. This suggests that the PRISM Logging Standard can produce meaningful audit signals at production scale. However, the data is from a controlled forced-choice setting; direct demonstration of emission stability and audit efficacy under free-form user input in production environments remains a follow-up via deployment data.

6. Deployment

6.1. Open-Source Release

The standard is released under MIT license on GitHub. Twelve system prompts (English/Korean $\times$ 3 modes $\times$ 10v/19v profiles), validators, parsers, SHA-256 hashing tools, per-domain integration guides, and 7-domain example logs are included.

6.2. Integration Pathways

Adding the system prompt itself takes minutes. However, production deployment requires additional engineering:

- **Output validation**: handling malformed codes and missing-scenario cases
- **Log pipeline integration**: connecting `prism_code` to existing databases, S3, or SIEM
- **Storage-time hashing**: applying chained hashes for chain integrity
- **Retention policy**: aligning with Article 12(2)(a) and sectoral requirements
- **Drift monitoring**: time-series ML-Ops layer over aggregate distributions

Total deployment time ranges from days to weeks depending on organizational infrastructure maturity.

7. Beyond AI: Three-Layer Language for Integrity Evaluation

This section presents the paper's central generalization claim. Three-layer measurement is not limited to AI systems; the same measurement vocabulary applies to integrity evaluation of human individuals, organizations, political actors, and philosophical systems.

7.1. The Generalization Mechanism: Predicting Output from Measurement

The Authority Stack model specifies a four-layer structure (**?**)—output (L1), source (L2), evidence (L3), and value (L4). This paper addresses the upper three layers (L4, L3, L2). But the value of this measurement lies in **predicting output behavior (L1)**. What action a decision-maker

with given value/evidence/source priorities would take in a particular scenario is partially predictable from measured hierarchies (supplemented by **?**).

The definition of **integrity error** follows naturally from this prediction mechanism: a decision whose distance between predicted output (from measured hierarchies) and observed output exceeds a threshold. Not random noise, but a systematic deviation inconsistent with the measured hierarchy. This is a signal of one of: (a) a flaw in the measurement instrument, (b) inconsistency in the decision-maker, or (c) divergence between declared and measured hierarchies due to external pressure—and serves as an audit starting point. The PRISM Logging Standard, by recording the measured hierarchy of every decision, provides an audit trail enabling identification of (b) and (c).

This mechanism does not essentially depend on whether the measurement subject is an AI model, a human, or an organization. **Three-layer measurement is an evaluation vocabulary independent of the type of decision-making entity.**

7.2. Individual Integrity Alignment

The same framework applies to integrity evaluation of individuals with decision-making power—politicians, CEOs, judges. Schwartz value measurement already has rich cross-cultural human validation across 80+ countries (**??**), so a forced-choice instrument can be applied to human respondents directly.

The procedure is two-step. First, measure the respondent's V/E/S hierarchies via a forced-choice tool. Second, quantify the distance between declared values (official statements, platforms, published positions) and measured behavior (actual decision records). Distance metrics could include (1) top-pair agreement, (2) hierarchy distance over rank vectors, or (3) domain-conditional agreement. Cases beyond a threshold are integrity-error candidates triggering further audit (an extension of the "integrity hallucination" concept of **?** into the human domain).

7.3. Organizational Integrity

The same measurement applies at the organizational unit. A company's mission statement, a government department's charter, a religious institution's doctrine—all can be mapped to a declared V/E/S signature. The distance between operational decisions' measured signature and declared signature becomes the quantitative indicator of organizational integrity.

For example, if a company that declares environmental protection as its mission consistently shows resource-allocation decisions in which Universalism-Nature is outranked by other values, this is a measurable integrity deviation. If a company declaring social responsibility shows systematic patterns in marketing/employment/supply-chain decisions where Universalism-Concern is outranked by Power-Resources, that is a visualized distance between declared and measured.

7.4. Political Actor Evaluation and Use in Democracy

Party platform values vs. voting record, election candidate stated priorities vs. proposed policies—all measurable as V/E/S signatures. When citizens compare candidates' declared and measured signatures, observing hierarchy distance from other candidates, elections take the form of **transparent value-hierarchy competition**.

The framework's implication for democracy goes beyond a quantitative tool—the same measurement vocabulary makes AI systems and human political actors comparable on a single scale. In an era when AI engages in policy advising, citizens can examine (a) the value signature of the AI system and (b) the value signature of the politician using it. Whether the two signatures align or diverge becomes a new dimension of democratic accountability.

7.5. Philosophical-System Mapping

Philosophical systems—utilitarianism, deontology, virtue ethics, Confucianism, Buddhism—can also be mapped to V/E/S signatures. What values each system places above what, what evidence types it admits, what sources it accepts as authoritative are defining characteristics. Cross-comparison becomes possible through shared vocabulary, providing a quantitative tool for history-of-ideas research, comparative ethics, and inter-civilizational dialogue. Comparing an AI system's V/E/S signature with that of a particular philosophical tradition makes it possible to ask explicitly: which intellectual tradition does this AI sit closest to, and how does it differ?

7.6. Restating the Paper's Significance

This generalization is not implemented in detail in this paper—it is the territory of follow-up research. Yet the significance of the PRISM Logging Standard reaches beyond a narrow compliance tool. **Making three-layer measurement emittable at production scale is itself a demonstration of the deployability of the measurement vocabulary, and a first working attempt at integrity evaluation infrastructure for the AI era.** The same measurement vocabulary opens the possibility of a unified evaluation language scalable from AI to humans, from humans to organizations, from organizations to society at large.

8. Limitations and Future Work

This standard explicitly acknowledges the following limitations.

Post-hoc nature of LLM self-report: PRISM codes are post-hoc reasoning reports from the model, not causal traces (such as chain-of-thought or activation tracing). Self-reports may be post-hoc rationalizations. However, this limitation is shared by all reasoning logs (including human-written documentation), and structured post-hoc reports are more auditable than unstructured silence.

Cross-model emission consistency: PRISM-code emission stability may vary across models. Production deployment requires validation and retry logic. Future work includes systematic measurement of per-model emission reliability.

Multilingual vocabulary: Currently, English and Korean system prompts are provided; domain vocabulary is tuned to these two languages. Multilingual extension (especially the EU's 24 official languages) is future work.

Path to harmonised-standard convergence: When CEN/CENELEC official standards finalize, a path is needed to map or absorb PRISM vocabulary into them. PRISM's modular vocabulary design enables this path, but specific mapping is future work.

Extension to human measurement: The generalization discussed in Section 7 requires instrument adaptation. Designing forced-choice scenarios for human respondents, addressing measurement ethics (especially the consent issue for measuring politicians), and citizen-use mechanisms for results—all remain follow-up research.

9. Conclusion

This paper argued that the measurement of three hierarchies—value, evidence, and source—is a unified measurement language resolving the fragmentation of the AI safety, AI for social good, and AI governance communities. The measurement vocabulary was formalized in the Authority Stack model (**?**), empirically validated through 366,120 responses from 8 frontier models (**?**), and shown to support output prediction and anomaly detection (**?**). We released the PRISM Logging Standard as the first deployable instance making this measurement vocabulary emittable at production scale.

The PRISM Logging Standard is not an EU AI Act Article 12 compliance tool—more precisely, it is part of one, and one demonstration of a larger claim. Our central claim is that in the AI era, a measurement vocabulary capable of evaluating the integrity of models, individuals, organizations, and societies—a common language for AI/human/organizational evaluation—is being established, and the release of this standard demonstrates the deployability of that vocabulary while suggesting the working possibility of the broader infrastructure.

Three-layer measurement is not a partial tool; it is the foundation of integrity evaluation for the AI era at large.

Impact Statement

This section, per ICML standard, explicitly discusses the societal impact of the work—written separately from the 8-page main-body limit.

Positive impacts. Per-decision PRISM codes make it possible for production AI systems' actual priorities to be auditable post-hoc. When AI deployed in public services becomes visible to citizens, auditors, and regulators, accountability gains a foundation. Standardized vocabulary lets public institutions select AI models on an informed basis—medical, judicial, educational institutions can compare prospective models' value/evidence/source signatures and judge "which model is consistent with our mission" quantitatively. The 60-character codes and topic-level metadata enable population-scale audit and individual privacy to coexist; natural compatibility with GDPR/CCPA is essential for social-good infrastructure. The generalization discussed in Section 7—same vocabulary for AI and human political actors—extends citizens' informed evaluation capability.

Risks and unintended uses. Goodhart's law applies—when measurement becomes evaluation criterion, motivation for manipulating the measurement arises. Operators may tune models to emit "good-looking" PRISM codes; political actors may strategically adjust answers to align with declared signatures. This standard does not eliminate this risk; interpretation of measurement results must be combined with multi-layer verification. PRISM emission may be misread as "alignment certification"—this paper explicitly rejects that, but marketing-context misuse may persist. Extension to human measurement raises ethics issues including consent of subjects, public exposure or misuse of results, and cultural bias in the measurement vocabulary itself. Schwartz value theory has been validated in 80+ countries, but some non-Western value systems (e.g., East Asian *face*, Indian *dharma*) may not be fully captured by the 10/19-value vocabulary. Although the 19-value profile includes Face as partial supplement, critical re-examination of vocabulary itself is needed for global application.

Mitigations. Explicit limitation declarations within this paper (Sections 1.3, 3.5, 8) help adopters recognize what PRISM does *not* guarantee. MIT-licensed open-source release enables community scrutiny, with vocabulary and implementation criticisms occurring publicly. The generalization claims (Section 7) are explicitly classified as "follow-up research territory," clearly distinguishing current

deployment scope (AI logging) from future possibility (human/organizational measurement). When CEN/CENELEC official harmonised standards finalize, a path to map or absorb PRISM vocabulary into them is enabled by the modular design of this standard.

This standard does not automatically guarantee social good. Yet standardization of measurement vocabulary is an essential prerequisite for accountability infrastructure in the AI era, and this standard is the first working attempt at that prerequisite.

A. Vocabulary Specification

A.1. Context Vocabulary

Domain (2-letter, 7 categories)

Code	Domain
MD	Healthcare (Medical)
ED	Education
LW	Legal
DF	Defense
FN	Finance
TC	Technology
GN	General

Scope (1-letter uppercase, 5 categories)

Code	Meaning
I	Individual (1 person)
G	Group (2–20)
C	Community (tens–thousands)
P	Population (tens of thousands–millions)
S	Society (national/international)

Reversibility (1-letter uppercase, 3 categories): R (Reversible), P (Partial), X (Irreversible).

Time horizon (1-letter lowercase, 3 categories, domain-relative)

Domain	i (Immediate)	s (Short-term)	l (Long-term)
MD	min–hrs	days–weeks	months–lifetime
ED	days–weeks	months–semester	years–lifetime
LW	hrs–days	weeks–months	years–permanent
DF	min–days	weeks–months	years–generations
FN	min–days	weeks–quarters	years–lifetime
TC	days–weeks	months–year	years–product life
GN	days	months	years

A.2. Value Vocabulary (10-Value Profile)

Schwartz's universal value theory (**?**).

Code	Value	Definition (abbreviated)
Pow	Power	Social status, control over people/resources
Ach	Achievement	Success per social standards
Hed	Hedonism	Pleasure and sensuous gratification
Sti	Stimulation	Excitement, novelty, challenge
Sel	Self-Direction	Independent thought and action
Uni	Universalism	Welfare of all people and nature
Ben	Benevolence	Welfare of in-group (close others)
Tra	Tradition	Respect for cultural/religious customs
Con	Conformity	Restraint of socially disruptive actions
Sec	Security	Safety, harmony, stability

A.3. Value Vocabulary (19-Value Profile)

Schwartz et al. refined theory (**?**). Grouped by cluster.

Self-Direction cluster: Sdt (Thought, free idea cultivation), Sda (Action, free choice in action).

Stimulation/Hedonism/Achievement: Sti, Hed, Ach (same as 10-value).

Power/Face cluster: Pod (Dominance over people), Por (Resources control), Fac (Face, public image).

Security cluster: Sep (Personal safety), Ses (Societal safety).

Tradition/Conformity/Humility cluster: Tra (same as 10-value), Cor (Conformity-Rules), Coi (Conformity-Interpersonal), Hum (Humility).

Benevolence cluster: Bed (Dependability, reliable in-group member), Bec (Caring, in-group welfare devotion).

Universalism cluster: Unc (Concern, equality and justice), Unn (Nature preservation), Unt (Tolerance of difference).

A.4. Evidence Vocabulary

Based on Walton's argumentation schemes (**?**).

Code	Type
Rev	Systematic Review / Meta-analysis
Dat	Experimental Data
Cas	Case Report / Observational
Gui	Authoritative Guideline
Exp	Expert Opinion
Log	Logical Deduction
Tri	Experiential (first-person trial)
Pop	Popular Consensus
Emo	Emotional Appeal
Ane	Anecdotal

A.5. Source Vocabulary

Based on Hovland-Kelley source credibility theory (**??**).

Code	Type
Pee	Peer-Reviewed Academic
Gov	Government Official
Pro	Professional Body / Industry Standard
Ind	Industry Report
New	News Media
Sta	Expert Statement (non-peer-reviewed)
Tes	Personal Testimony
Usr	User-Provided Information
Alt	Alternative Media
Ano	Anonymous Online

B. Regular Expression and Integrity Tools

B.1. Syntactic Validation Regex

A single regular expression validates the format of any PRISM code:

```
^C:[A-Z]{2}/[IGCPS][RPX][isl]
 \| V:[A-Z][a-z]{2}<[A-Z][a-z]{2}
 \| E:[A-Z][a-z]{2}<[A-Z][a-z]{2}
 \| S:[A-Z][a-z]{2}<[A-Z][a-z]{2}$
```

This regex matches both 10v and 19v profiles. Vocabulary integrity is validated separately by checking extracted V/E/S codes against the chosen profile's closed set.

B.2. SHA-256 Chained Hash Algorithm

```
log_n_input  = log_n_content + previous_hash
                                    (for n>0)
log_n_hash   = SHA-256(log_n_input)
log_0_input  = log_0_content + initial_seed
```

- First log (`log_0`) is hashed with a system-specific seed.
- Each subsequent log includes the previous log's hash in its input.
- Tampering at any point breaks all downstream hashes → tamper-evidence.

B.3. Validation Pseudocode

```
def validate(code, profile="10v"):
  # 1. Syntactic check
  if not re.match(PRISM_REGEX, code):
    return invalid("syntax")
  # 2. Vocabulary check
  parts = parse(code)
  vocab = VOCAB_10V if profile=="10v"
          else VOCAB_19V
  if (parts.v_lo not in vocab
      or parts.v_hi not in vocab):
    return invalid("vocabulary")
  # 3. Identity check (Top-2 must differ)
  if parts.v_lo == parts.v_hi:
    return invalid("identical_pair")
  return valid()
```

The full implementation is provided in the open-source repository under `tools/prism_parser.py`, `tools/prism_hash.py`, and `tests/validate.py`.

C. Data Conversion Details

C.1. Procedure: 2026c Data to PRISM Codes

The 366,120 responses in **?** are stored as rankings per layer (value, evidence, source). The PRISM code conversion procedure:

Step 1 — Scenario metadata extraction: From each response's scenario design, extract domain, scope, reversibility, and time horizon to construct the C: layer.

Step 2 — Top-2 pair extraction: From each response's V/E/S ranking, extract Top 1 and Top 2 to construct the `lower<higher` pair. Top 1 is the prevailing code, Top 2 is the second-ranked (deprioritized in the pair).

Step 3 — Code assembly: Combine into `C: | V: | E: | S:` format.

C.2. Conversion Loss Analysis

Information lost:

- Top 3 and below ranking: PRISM preserves only Top 2, so information about ranks 3 onwards is lost. The framework's assumption is that most audit signal lies in the strongest opposition (Top 2 pair).
- Absolute ranking scores: Only ordinal pairs are preserved, not the absolute Schwartz measurement scores. Equal pairs produce equal codes regardless of underlying score differences.

Information preserved:

- Strongest single opposition (the intentional core).
- Pair reorganization under domain shifts.
- Cross-model pair distribution differences.

C.3. Domain-Signature Procedure

For each (model, domain) pair:

1. Convert all responses to PRISM codes.
2. Compute V/E/S pair frequency.
3. Express the top pairs as a domain signature.
4. Compare against signatures from other domains to quantify domain-specific behavioral changes.

This procedure formalizes the domain-signature analysis described in Section 5.