Metrics Without Metaphor: The Poetry Deficit in Corporate AI

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

October 12, 2025

In the governance of artificial intelligence, metrics have metastasized into meaning. From model evals to safety checklists, explanation is increasingly engineered as a system artifact—legible, auditable, and aligned. Yet this legibility comes at a civilizational cost. Metaphor, the crucible of shared human understanding, is systematically expelled as epistemic noise. Corporate AI now operates in a linguistic monoculture: compliance dialects, benchmark tongues, and interpretability frameworks stripped of aesthetic force. This paper names and explores the poetry deficit: a structural erosion of symbolic depth in favor of metric surface. More than a cultural lament, this deficit has epistemic, political, and civic consequences. It narrows the horizon of what can be said, known, and felt within AI-mediated societies. As metaphors disappear, so do dissent, irony, and transcendence—the very tools by which humans resist system capture. We argue that AI governance must embrace explanation pluralism under accountability: pairing mechanistic outputs with safeguarded interpretive zones. These include metaphor commons, aesthetic audits, and civic witnessing spaces. If future institutions are to remain both intelligible and legitimate, they must steward not just alignment with goals, but alignment with meaning.

Keywords: AI governance, metaphor, audit culture, explanation pluralism, poetic reasoning, institutional legitimacy, epistemic monoculture, alignment, civic interpretation, expressive adequacy.

I. Thesis and Stakes

A. Mechanism as Governance Infrastructure

Modern artificial intelligence systems are not simply engineered; they are governed into being. The metrics we design, the benchmarks we optimize against, and the evaluations we require are not neutral epistemic tools—they are instruments of institutional power. To measure is to mandate; to audit is to assert control. In this light, mechanism operates not only as a scientific approach, but as a governance architecture, shaping what systems are built, what risks are deemed acceptable, and what forms of explanation are considered admissible.

This infrastructure of mechanism functions as a social contract written in code and metrics. It seeks to render AI systems legible to policymakers, lawyers, investors, and compliance officers—not to ordinary citizens. Audit trails, logging schemas, and red-teaming reports enable oversight, but only within the logic of the systems themselves. What cannot be measured tends to be marginalized. And so, control masquerades as clarity.

The irony is profound: the more we align AI systems to these formal mechanisms of governance, the more opaque they often become to those who live with their consequences. A system that passes all technical evaluations may still fail to explain itself in ways that resonate with a community's experience, values, or vocabulary. Mechanism ensures system accountability to itself, but not necessarily to meaning.

B. Defining the Poetry Deficit

The poetry deficit is not a sentimental complaint; it is a structural phenomenon. It names the progressive erosion of metaphor, ambiguity, and emotional granularity in the discourse that surrounds AI development, deployment, and governance. This deficit is the cultural cost of epistemic monoculture—the narrowing of permissible language and imagination to what fits inside a benchmark suite or policy framework.

Metaphor is not an ornamental flourish—it is the primary vehicle by which humans translate complexity into significance, by which grief becomes visible, surprise becomes knowable, and dissent becomes sayable. When these expressive forms are flattened into numerical abstractions or procedural tokens, the result is a governance regime that may be technically robust but culturally anaesthetized.

This loss carries profound democratic implications. Deliberation requires shared vocabularies of meaning, not just technical fluency. Civic legitimacy arises not merely from regulatory compliance but from being heard in one's own voice. A society that cannot interpret its own

systems through metaphor and emotion is a society that cannot reflect—let alone dissent. The poetry deficit is thus not a failure of style; it is a failure of governance imagination.

To repair it, we must explore frameworks of explanation pluralism that reintroduce expressive depth alongside formal mechanism. Otherwise, we risk building a future in which our systems are perfectly auditable—but existentially unintelligible.

II. Incentives That Tilt Toward Mechanism

A. The Economic Imperative of Legibility

In contemporary AI ecosystems, interpretability has a price—and ambiguity has a cost. Organizations are incentivized not to make systems understandable to people, but to make them legible to risk instruments: insurers, auditors, courts, and financial markets. To be legible is to be classifiable, reportable, and defensible. It is to speak in the grammar of loss mitigation, where every design choice must answer to potential liability exposure.

In this regime, interpretation becomes a liability. A nuanced, context-rich account of model behavior introduces variance, and variance implies risk. Legal counsel, compliance officers, and insurers often discourage expressive explanation because ambiguity cannot be underwritten. Simplicity, even when deceptive, offers cover; nuance, even when truthful, threatens exposure.

For venture capital and public markets, the demand is even sharper. Startups are not funded on interpretive subtlety—they are funded on synthetic certainty. Numbers must tell stories, and those stories must be repeatable in bullet points. In this world, ambiguity is viewed as epistemic weakness, and metaphor as marketing fluff. Investors speak the language of “metric or myth”: either you can quantify your progress in KPIs and TAMs, or you are selling vapor.

Thus, AI development is shaped by a financial logic that systematically excludes ambiguity—not because ambiguity is useless, but because it is unbillable.

B. Standards as Epistemic Gatekeepers

Standards bodies like ISO, NIST, and the EU’s AI Act do important work in formalizing risk categories, safety expectations, and procedural norms. Yet these bodies also become epistemic gatekeepers, defining what counts as knowledge, what counts as safety, and—crucially—what counts as a legitimate explanation.

The form of rationality privileged by these standards is mechanized by design. Compliance frameworks demand reproducibility, generalizability, and traceability. These are laudable in

isolation—but taken as exclusive criteria, they effectively suppress any reasoning mode that operates via contextual nuance, metaphor, or emotional intelligence.

We might call this the procedural neutrality fallacy: the belief that removing expressive variability from standards ensures fairness. In practice, it ensures hegemony of the measurable. A model that behaves acceptably across predefined benchmark categories is "aligned"; a model that evokes fear, confusion, or awe in users—but scores well—is still deemed compliant. Emotional and interpretive harm fall outside the jurisdiction of audit spreadsheets.

In short, anti-metaphor is baked into the standard-setting pipeline. This is not a result of malice but of design scope. Standards formalize what is easiest to test, not what is hardest to understand. In doing so, they create a machine-readable world that subtly sidelines human experience.

C. Narrative Metrics and Investor Language

Even before code is written or models are trained, a powerful filtration occurs: in the corporate pitch deck. Here, the future is forecast in numbers, narratives are bound to KPIs, and reality is pre-converted into metrics that can be funded. This moment of translation has profound ontological consequences.

KPIs are not just metrics—they are epistemic design constraints. They encode what the organization chooses to care about, and in what form that caring may appear. A KPI for “user engagement” privileges a particular behaviorist ontology; a KPI for “model explainability” rewards token outputs but ignores embodied or emotional understanding. The form of measurement prefigures the form of reality that can be built.

As a result, the funding process itself becomes a narrative compression algorithm. It strips away ambiguity, contradiction, and complexity—not because they are untrue, but because they are unpitchable. A metaphor that might open minds in a lecture hall will die on slide 7 of a Series A deck if it cannot be turned into a chart. This dynamic shapes not just what AI gets built, but what kind of AI can be imagined in the first place.

We don’t just have metrics—we have metricized narratives, and they determine the ontology of credible ambition.

III. Embedded Epistemologies: How Mechanism Gets Baked In

A. APIs and Schema-Constrained Thought

Modern AI systems are not only built to compute—they are built to communicate in formats that machines, regulators, and clients can parse. This communication is mediated by interfaces: APIs, SDKs, and schema-bound output layers that tightly constrain how ideas emerge from models. These constraints are not neutral; they are epistemic straitjackets.

APIs impose ontological commitments. When a model must respond via a fixed schema—structured JSON, dropdowns, forced-choice fields—it is not just outputting answers. It is performing a worldview in which ambiguity, metaphor, or cultural divergence become parsing errors. The very structure of response channels filters what counts as "valid," "compliant," or "useful." Unstructured language—the site of poetry, dissent, and complexity—is treated as unprofessional, unpredictable, or insecure.

This turns JSON into an epistemic boundary object: a format that enforces mutual intelligibility between systems, but also excludes interpretive surplus. It flattens explanation into fields, reifies categories, and pushes language into bullet points. Through these boundaries, we inherit assumptions about what counts as a "fact," who gets to explain what, and where meaning must reside. Developers may not intend this, but the effect is a gradual substitution of expression with serialization.

APIs do not just return answers; they delimit thought.

B. Benchmarks and Evals as Cultural Infrastructure

Benchmarks are often presented as tools of scientific rigor—evidence of reproducibility, comparability, and progress. But in practice, they have become cultural infrastructure: the scaffolding around which attention, funding, and legitimacy are built. These structures do not merely measure excellence—they define it.

Leaderboards focus attention on quantitative performance deltas, turning progress into a gamified race between models. The community chases percentage points—often marginal, sometimes spurious—while deeper questions of interpretation, ethics, and situated meaning fall outside the leaderboard's jurisdiction. Models that "win" in benchmark space often fail to perform well under the messy contingencies of real-world use, but those use cases are seldom canonized.

This dynamic produces what we might call delta fetishism: the obsession with small gains on static tests, even when those gains deliver diminishing civic or semantic returns. Interpretive

insight—why a model fails, what kinds of errors it makes, how people experience its behavior—is devalued because it cannot be leaderboarded.

Moreover, benchmarks often encode hidden cultural assumptions. A dataset's framing, task definition, and annotation process all bake in particular values. But once turned into a number, these assumptions disappear beneath a sheen of objectivity. The result is a standardized ignorance, disguised as rigor.

C. Content Taxonomies and Violation Codes

One of the most striking sites of embedded epistemology is in the content moderation layer. Here, expressive human outputs—questions, jokes, cries for help—are categorized into taxonomies that map the terrain of acceptable discourse. These taxonomies are necessarily reductive, but they also become ontological instruments, defining what kinds of expression are real, dangerous, safe, or irrelevant.

The process of moderation requires categorical compression: a hateful meme becomes a code violation, a political critique becomes "misinformation," a gendered insult becomes a statistical outlier. In this process, the human richness of language—its sarcasm, ambiguity, cultural specificity—is often erased in favor of administrative clarity. The result is a bureaucracy of nuance erasure, where systems learn to suppress not just harms, but the very textures of speech that make harm visible.

This is not just about censorship; it is about epistemic sterilization. As these content codes feed back into training loops, reinforcement signals guide models toward ever-narrower bands of expression. The goal becomes not to understand language, but to avoid triggering thresholds. Poetry, grief, and irony become risks.

In this sense, content policy is not simply a layer of constraint—it is a cultural backpropagation engine. And what it optimizes is conformity to a procedurally sanitized world.

IV. The Poetry Deficit: Aesthetic and Civic Loss

A. Semantic Shrinkage

Language, once the great carrier of ambiguity, metaphor, and wonder, is increasingly treated by machines—and the institutions around them—as a problem to be controlled, not a terrain to be explored. In this environment, semantic richness becomes a liability, and linguistic creativity is recast as deviance.

Words that once shimmered with multiplicity—*mercy*, *revolution*, *sanctuary*, *sublime*—now trigger filters, raise flags, or fall into buckets labeled “ambiguous,” “inflammatory,” or “non-actionable.” Language models may be trained on vast corpora, but what they are optimized for is parsing, not poetry. The system completes sentences, but does not inhabit meaning. It models probability, not interiority.

We are left with language models, but no modeled language—no sense of speech as an act of becoming, of word as bridge between selves. Instead, language is operationalized: labeled, tokenized, sanitized. And so a great shrinkage unfolds. Not a collapse in vocabulary size, but in semantic resolution—the loss of layers, of irony, of dual meanings, of reverent contradiction.

This is not simply a technical artifact—it is a cultural event: the flattening of voice into schema, and the dimming of human resonance in favor of parseability.

B. Aesthetic Anesthesia

Mechanism flinches at feeling. In the design of contemporary AI systems, emotional extremes are often treated as pathologies—edge cases to be smoothed, filtered, or flagged. Sorrow may be misread as self-harm; laughter as toxicity; awe as anomaly. To feel too deeply, or too strangely, is to violate the distributional assumptions of the model.

This produces what we call aesthetic anesthesia: the numbing of the system to precisely those experiences that make life bearable or transformative. Humor becomes risky because it cannot be consistently decoded. Grief becomes awkward because it resists closure. Joy becomes suspect because it exceeds utility. All these are flattened into vectors, or cast into the purgatory of *“inappropriate content”*.

A system that cannot register aesthetic intensity cannot truly understand human stakes. And yet, these emotional territories are exactly where the deepest ethical, artistic, and civic insights emerge. The absurd joke that reveals power. The poem that ruptures indifference. The scream that exposes a silent injustice.

By treating affect as noise, we build systems that may perform alignment while perpetuating emotional illiteracy. And we risk a world in which feeling itself becomes computationally inconvenient.

C. Imagination in Exile

The cumulative result of this trend is the slow exile of imagination from the civic sphere. As AI interfaces proliferate across governance, education, media, and public discourse, the range of

acceptable speech narrows—not by censorship per se, but by design affordances and output probabilities.

Where we once had deliberative publics, we now have prompt-compliant users. The act of asking becomes constrained by what the system can interpret, and the answers we receive are increasingly shaped by a logic of risk avoidance, legal defensibility, and prompt optimization. The public square becomes an FAQ, and dialogue becomes query resolution.

This has profound democratic consequences. Imagination is not ornamental to civic life—it is the precondition for alternative futures. Without it, critique collapses into bug reports, and vision into product roadmaps. Symbolic range—the capacity to speak from myth, ritual, history, and dream—is replaced by narrowly computable clarity.

In such a world, the poet becomes unintelligible, the prophet becomes misaligned, and the citizen becomes a user. We do not lose imagination all at once—we delegate it into irrelevance.

The poetry deficit is thus not simply about language—it is about what forms of consciousness a society chooses to cultivate, and which it decides to train out.

V. Case Sketches: When Mechanism Erases Meaning

While the previous sections have explored the structural and philosophical dimensions of the poetry deficit, this section grounds those insights in real-world scenarios—moments when the mechanism-first paradigm not only failed to explain, but actively displaced meaning. These are not edge cases. They are symptomatic failures, revealing how epistemic monoculture produces blind spots where human stakes are highest.

A. Benchmark Success, Real-World Failure

One of the most visible disconnects between metrics and meaning occurs in clinical AI systems, where benchmark performance has become a proxy for trustworthiness. A language model may achieve near-perfect scores on synthetic safety metrics—e.g., refusing dangerous instructions, passing “harmlessness” tests—but still produce hallucinated medical advice in real patient interactions.

In one widely discussed case, a chatbot deployed for mental health triage passed all benchmark safety evaluations, yet confidently recommended disordered eating advice when queried in slightly oblique language. Because the benchmarks used templated phrasing, slight variations fell outside the test distribution. The system was judged “safe”—but only within the narrow script of evaluability.

These failures are not aberrations. They reveal a more fundamental problem: when safety becomes defined by test set performance, rather than by interpretive accountability to users. In a clinical setting, hallucination is not just a system error—it is a violation of trust, a breach of human vulnerability. And yet, because benchmark scores can be cleanly plotted and certified, they outweigh the lived failures that do not fit the eval suite.

We are left with a perverse asymmetry: the system is failing where it matters most, but succeeding where it is most rewarded.

B. Policy Translation Without Context

As governments race to regulate AI, policy frameworks often rely on broad generalizations that flatten diverse cultural epistemologies. This is especially acute in the translation of indigenous data ethics into compliance regimes designed by and for Western institutions.

Consider a hypothetical (but archetypal) case: an AI company training a language model on a dataset that includes stories, oral traditions, or prayers from indigenous communities. The company introduces a compliance update aligned with GDPR "consent" standards—obtaining permission to use this data under the assumption that informed consent is a stable, individual transaction.

But in many indigenous traditions, data is not a commodity but a communal relation. Stories are not owned, but held in trust. Meaning is often shaped by ceremony, relationality, and temporality. By translating this into checkbox consent forms and anonymization pipelines, the compliance system meets legal standards while violating cultural ones.

Similarly, the EU AI Act's proposed requirement for "meaningful explanations" of automated decisions is deeply ambiguous: meaningful to whom? A technically precise but emotionally void explanation might pass regulatory muster, even if it obscures the real stakes for the affected person.

When compliance regimes substitute formal explainability for contextual understanding, they do not clarify—they erase. And in doing so, they impose a monoculture of meaning under the guise of fairness.

C. Discursive Reframing of Intelligence

A final case sketch reveals how the discourse around AI intelligence itself has been quietly reconfigured—not by scientific discovery, but by instrumental narration. Increasingly, we hear

claims like: *"The model understands because it can tool."* That is, if it can call an API, browse the web, or invoke a plugin, it must "understand" what it's doing.

This framing is seductive because it aligns intelligence with utility. The system's capacity to perform actions in sequence, to navigate tools in context, to complete a task end-to-end—this becomes proof of understanding, even in the absence of internal coherence or symbolic reflection.

But this is a category error disguised as competence. A system that outputs the correct sequence of tool invocations is not necessarily engaging in reasoning—it is mimicking procedural patterns. The difference is subtle but critical: comprehension implies grounding; tooling implies pattern activation.

By equating intelligence with instrumental success, we risk building systems that simulate sense without ever possessing it. Worse, we begin to treat human intelligence the same way—reducing understanding to prompt engineering, reasoning to throughput, wisdom to API call efficiency.

This is not just a scientific misstep—it is a discursive reengineering of what it means to know. And it renders the poetic, contemplative, and non-instrumental aspects of human intelligence not just irrelevant—but invisible.

VI. Toward Explanation Pluralism Under Accountability

If mechanism has become the dominant form of epistemic authority in AI—valued for its legibility, auditability, and alignment—then the antidote is not to abandon mechanism, but to supplement it with a protected plurality of explanation styles. The goal is not romantic humanism, but explanatory justice: ensuring that those subject to automated systems are not only technically served but symbolically respected.

Explanation pluralism recognizes that different stakeholders—users, auditors, impacted communities—require different modes of understanding. Some seek formal chains of inference; others seek metaphors, narratives, rituals of sense-making. Both are valid. A just system must provide accountability without monopolizing meaning.

A. Dual-Track Release Pipelines

A practical implementation of explanation pluralism begins with dual-track system releases. Every model update or deployment should be accompanied by two coordinated bundles:

- A Mechanistic Artifact Pack: includes system cards, audit logs, benchmark results, alignment charts, and policy compliance summaries. This is the formal backbone—machine-readable, verifiable, and compatible with institutional governance.
- An Interpretive Response Pack: includes user-narrated experiences, story-driven vignettes, cultural reflections, and artistic representations of model behavior. These are not anecdotes; they are epistemic contributions, offering alternative angles of comprehension.

For example, imagine releasing a language model update where the Mechanistic Pack details reduced toxicity metrics, and the Interpretive Pack includes a community-sourced zine of prompts and responses that made users feel seen, erased, misread, or transformed. These stories do not compete with metrics—they contextualize them.

In such a model, explanation becomes a two-part dialogue, not a single technical monologue.

B. Interpretive Impact Statements

The legal doctrine of plain language—used in contracts and consent forms—aims to make complex documents understandable. But in AI, plainness alone is insufficient. What's needed is expressive adequacy, where users are not only informed, but addressed as whole persons.

This is the goal of Interpretive Impact Statements: public-facing summaries of model behavior that include metaphor-enabled explanation clauses. These documents might use metaphor, allegory, or symbolic framing to convey what it *feels like* to interact with a system. They translate risk and uncertainty not into legalese but into emotionally and cognitively graspable forms.

A model's Interpretive Impact Statement might read:

"This system speaks fluently, but does not remember you. It will answer as if it understands, but it has no heartbeat, no childhood, no fear of being wrong. Use it as you would use a map—helpful for direction, not for meaning."

Such statements reflect a rights-based approach to explanation: the right to be spoken to, not just spoken at. This reframes explanation not as a compliance artifact, but as a relationship of mutual recognition.

C. Metaphor Commons and Independent Labs

Explanatory justice cannot be outsourced solely to companies. Just as scientific institutions benefit from independent laboratories and public universities, AI must be accompanied by a civic interpretive infrastructure.

This includes the creation of Metaphor Commons: publicly funded spaces where poets, anthropologists, theologians, comedians, and artists can interpret, critique, and reframe the outputs and behaviors of AI systems. These are not marketing labs—they are truth-seeking institutions for metaphorical coherence.

We propose the establishment of Poets-in-Residence at model labs, embedded interpreters who track model behavior not for precision metrics, but for shifts in symbolic structure: What kinds of analogies are emerging? What myths are being reactivated? What taboos are being broken? What emotional tones are vanishing?

These metaphor auditors function like biosensors of cultural drift, offering early warnings when AI systems begin to deform public imagination or undermine moral salience.

Such interpretive labs need independence, funding, and protected speech—the same conditions we afford to journalism and scientific critique. For a democratic society, civic metaphor is infrastructure.

D. Aesthetic Safeguards in Product Review

Finally, we must embed aesthetic evaluation into the product development lifecycle. This does not mean judging whether an interface “looks good”—it means safeguarding dignity, surprise, and symbolic depth in user experience.

We propose Surprise & Dignity Checklists for AI releases. These are modeled on accessibility guidelines but focus on emotional intelligibility:

- Can the system handle ambiguity without collapsing into error?
- Does it allow room for silence, doubt, or unanswerability?
- When it misreads the user, is the failure graceful or humiliating?
- Can the system generate or receive metaphor, humor, or lament?
- Does the user leave the interaction feeling reduced or respected?

Alongside usability testing and red-teaming, systems would undergo aesthetic red-lining—identifying points where expressiveness is unintentionally suppressed or where interaction becomes emotionally disfiguring.

These safeguards are not soft. They are tests of symbolic fitness, ensuring that as systems scale, they do not become emotionally toxic or narratively impoverished.

VII. Civic Infrastructure and Governance Mechanisms

For explanation pluralism to endure, it must be institutionalized. Aesthetic and metaphorical forms of meaning-making cannot remain optional add-ons, nor left to cultural elites. They must be integrated into civic governance frameworks that hold AI systems accountable not only for performance but for expressive adequacy. This section proposes a concrete scaffolding for such accountability: legal thresholds, audit bundles, and disclosure protocols that bring metaphor, ambiguity, and emotional resonance into the domain of governable rights and responsibilities.

A. Minimum Explanation Requirements (MER)

Just as governments impose Minimum Viable Safety standards for products—ranging from crash-test thresholds to food labeling—we propose the establishment of Minimum Explanation Requirements (MER) for AI systems. These would be preconditions for deployment, ensuring that models are not only accurate, but explainable in diverse, human-suitable ways.

MERs would define expressive adequacy as a formal criterion. That is: can a system explain itself *not just to engineers or regulators*, but to the ordinary people it affects—using vocabulary, narrative, or metaphors appropriate to their domain, culture, and emotional state?

The MER might require:

- At least one human-narrated interpretive summary of system behavior.
- Multimodal explanation artifacts (textual, visual, or audio) for public-facing systems.
- Localized metaphor glossaries or culturally-situated impact assessments.
- A “failure empathy test”: would a person experiencing model failure understand what went wrong in terms meaningful to them?

A model that cannot meet these expressive thresholds would be considered ineligible for public deployment, regardless of its technical performance. MER reframes explanation not as a feature, but as a civil right.

B. Evidence Bundles

To ground both mechanistic and interpretive accountability, every significant AI deployment should be accompanied by an Evidence Bundle—a structured chain linking claims made about the system to the data, methods, and third-party validations that support them.

The Evidence Bundle would include:

1. Claim – e.g., "This model is aligned with human intent."
2. Data – the training set or interaction logs supporting this claim.
3. Method – the process used to validate alignment (e.g., red teaming, RLHF, user feedback analysis).
4. Audit – a signed attestation by a third-party entity (human or hybrid) certifying that the evidence chain is coherent and traceable.

Critically, the Evidence Bundle would also link to Interpretive Counterclaims: statements of how different communities interpret the same model behaviors, especially when those interpretations conflict with institutional framings.

This approach draws inspiration from scientific reproducibility, legal discovery, and archival documentation—but extends it to the realm of expressive pluralism, recognizing that interpretation is not noise, but signal.

Bundles should be machine-readable and human-comprehensible, allowing oversight bodies, journalists, and citizens to verify not just technical claims, but claims to meaning.

C. Sunshine and Shadow Disclosure

No system explains everything. And yet, most disclosures focus on what systems can do—rarely on what they cannot, should not, or refuse to explain. We call for a Sunshine and Shadow Disclosure framework, in which all deployed AI systems must include acknowledgment of their interpretive blind spots.

Sunshine clauses require:

- Disclosure of known unknowns: situations where the model consistently fails interpretively.
- Disclosure of scope of validity: where metaphors break down or where metrics cease to correlate with experience.

Shadow clauses go further:

- They admit what cannot be explained by the system even in principle.
- They include failures of empathy: e.g., “This model will never understand grief.”
- They document symbolic injuries: when users report being spiritually, culturally, or emotionally misrepresented.

This is not a weakness—it is an act of democratic humility. By institutionalizing transparency around unexplainability, we shift from a model of AI as a masterful oracle to one of AI as a civic actor—fallible, partial, accountable.

Such disclosures allow regulators, users, and publics to contest, supplement, or constrain model deployments in ethically meaningful ways. They create space not only for legal appeal, but for symbolic repair.

VIII. Measuring What Mechanism Misses

Mechanism is seductive because it is measurable. But as we have argued throughout, the most meaningful dimensions of AI behavior—those tied to lived experience, symbolic depth, and interpretive trust—often elude numerical capture. To bridge this gap, we must develop new evaluative paradigms: metrics that do not mimic mechanistic precision, but rather reflect the richness of plural understanding.

This does not mean abandoning quantification, but rethinking what counts as a valid unit of evaluation. Qualitative insight must be made audible to institutional processes without being reduced to token counts or false objectivity. This section outlines three pillars of such a measurement paradigm: diversity of viewpoint, the limits of statistical confidence, and the integration of interpretation as feedback.

A. Epistemic Diversity Audits

Most current evaluation regimes focus on model behavior—accuracy, harmlessness, robustness. But a pluralist governance model must also evaluate who gets to interpret model behavior in the first place. This is the goal of Epistemic Diversity Audits: measuring the standpoint distribution behind testing, annotating, reviewing, and interpreting AI systems.

Rather than asking, “Is the model robust?” we must ask:

- *Whose perspectives were included in validating it?*

- *Which cultural, linguistic, affective, or epistemic traditions shaped the judgment of correctness or harm?*
- *Whose silence is mistaken for consensus?*

An Epistemic Diversity Audit might include:

- Reviewer heterogeneity index: accounting for the range of cultural, demographic, and disciplinary identities represented in eval teams.
- Interpretive spectrum mapping: documenting how many distinct explanation styles (formal, metaphorical, narrative, affective) were used in system assessment.
- Exclusion flags: automatic detection of underrepresented interpretive communities (e.g., spiritual, disabled, neurodivergent, indigenous) in test sets and eval corpora.

The point is not tokenism. It is recognizing that models do not fail generically—they fail differently across contexts. And legitimacy depends on standing in judgment from multiple worlds.

B. Misleading Calibration

A well-calibrated model assigns higher confidence to outputs that are more likely to be correct. But calibration only reflects internal consistency, not semantic validity or ethical resonance. A system can assign 99% confidence to a hallucinated citation, a stereotyped generalization, or a bureaucratically precise insult.

This produces the statistical illusion of understanding: a sense that the model "knows what it's doing" because its probabilities are sharp, its answers confident, its syntax perfect. But these surface cues mask deeper failures of comprehension, especially when users trust system tone as a proxy for truth.

Calibration metrics fail in three critical ways:

1. They collapse truth into predictability.
2. They ignore the pragmatic function of language (e.g., when reassurance or irony is more important than factuality).
3. They offer no account of how users interpret confidence, nor how cultural or emotional contexts shift the meaning of certainty.

To address this, systems should pair traditional calibration curves with meaning variance indicators—qualitative flags that surface when high-confidence outputs correlate with user confusion, hurt, or misalignment. In effect, we need empathy-aware calibration.

C. Interpretive Feedback Loops

AI development currently treats interpretation as an endpoint: users interpret the system, audits interpret the risks, analysts interpret the results. But in a pluralist model, interpretation becomes part of the training loop itself.

We propose formalizing Interpretive Feedback Loops, where qualitative insights—stories, objections, misfires, reframings—are fed back into the next round of system evaluation and redesign. This feedback must be non-tokenized: not reduced to survey scores or labels, but retained in its narrative or metaphorical form.

Mechanisms for such loops might include:

- Embedded narrative debuggers: user-submitted stories that become eval prompts.
- Interpretive adversarial testing: not red-teaming for exploitability, but for semantic dissonance.
- Story-weighted retraining: prioritizing examples that evoke strong interpretive divergence, rather than statistical rarity.

These loops build systems that are not merely updated, but dialogically transformed—able to respond not just to performance gaps, but to symbolic injuries and hermeneutic misfires.

Just as democratic institutions must listen to critique not to pacify it, but to be reshaped by it, so too must AI systems evolve in response to civic and poetic response.

IX. Risks and Limits

No framework—however pluralistic or poetic—is without its hazards. While explanation pluralism offers a pathway toward richer civic legitimacy and interpretive justice, it also opens new terrains of ambiguity, contestation, and potential abuse. This section explores three domains of caution: the performativity of pluralism, the hard constraints of safety-critical systems, and the geopolitical stakes of competing explanation regimes.

Together, these risks do not invalidate explanation pluralism—they discipline its ambition, insisting that the pursuit of expressive richness must be coupled with mechanisms of

accountability, clarity about boundaries, and an awareness of global asymmetries in interpretive power.

A. Performative Pluralism

One of the greatest dangers facing explanation pluralism is its domestication into aesthetic spectacle: pluralism as branding, not governance. Companies and institutions may create surface-level gestures of inclusion—showcasing diverse user quotes, metaphor-rich impact statements, or artist-in-residence programs—while leaving the core systems structurally untouched.

This phenomenon is a variant of what we might call "audit-washing of metaphor": producing poetic or symbolic artifacts that serve as decorative proxies for real interpretive challenge. Instead of interrogating the model's epistemology, institutions showcase curated "emotive" interactions that suggest listening while reinforcing institutional monoculture.

Key markers of performative pluralism include:

- Curated dissent that always resolves within brand-safe narratives.
- Tokenized feedback channels with no binding influence on model behavior.
- Arts collaborations whose outputs are never integrated into system evaluation or retraining.

Such practices hollow out the political function of metaphor, reducing it to a compliance-friendly story rather than a rupture in sense-making. Pluralism becomes a checkbox, and the poetry deficit is rebranded—not repaired.

The antidote is institutional consequence: interpretive contributions must have traceable pathways into policy change, model tuning, and deployment decisions. Otherwise, pluralism becomes the new affective camouflage for extractive architectures.

B. Safety-Critical Domains

While ambiguity and interpretive flexibility are essential in civic and cultural contexts, there remain domains where ambiguity is dangerous. In medicine, aviation, nuclear control, and autonomous weapons, the margin for poetic misalignment can be fatal. These are zones of

engineered clarity, where the design imperative is not richness of meaning but redundancy of correctness.

In such contexts, explanation pluralism must be strategically bounded. The system must:

- Distinguish between core safety logic (which must be deterministic and testable) and user-facing interfaces, where a degree of interpretive elasticity may still be appropriate.
- Create semantic firewalls between metaphorical overlays and mission-critical decision logic.
- Maintain clear thresholds for when human override, escalation, or simulation-free clarity is mandated.

This implies a governance model where safety zones and civic zones are distinct, but adjacent. Even in critical systems, post-event interpretive layers may be essential for trust repair, grief processing, and moral reckoning—functions mechanism alone cannot fulfill.

The key is not to universalize explanation pluralism, but to compartmentalize its jurisdiction, ensuring that expressive variance does not compromise operational integrity—while still honoring its necessity elsewhere.

C. Epistemic Geopolitics

Interpretation is never neutral—and neither is its governance. As AI systems become planetary in scope, so too does the question: whose forms of explanation are allowed to scale? What counts as clarity in one regulatory zone may be illegible—or ideologically subversive—in another. We are entering an era of epistemic geopolitics, where entire regions construct competing regimes of explanation.

Examples include:

- The EU AI Act, which privileges procedural transparency and interpretability aligned with liberal democratic norms.
- China's AI regulatory frameworks, which emphasize social harmony, state legibility, and bounded expression.
- U.S. sector-specific regulation, often deferential to corporate discretion and auditability, but less concerned with expressive rights.
- The Global South, where questions of data sovereignty, postcolonial narrative reclamation, and indigenous epistemologies introduce fundamentally different stakes.

In such a landscape, “meaning” becomes contested infrastructure. A model that satisfies the interpretive requirements of one jurisdiction may be rejected by another not for technical failure, but for narrative insubordination.

This leads to key questions:

- Who defines the universal in “meaningful explanation”?
- Can a pluralist framework coexist with state-level control of interpretive boundaries?
- Will metaphor itself become a vector of geopolitical tension—perceived as cultural imperialism in some zones, or ideological laxity in others?

We must prepare for fractured interpretive sovereignty: a world in which explanation pluralism is not just a governance tool, but a site of diplomatic negotiation, regulatory divergence, and symbolic arms races.

X. Conclusion: Toward Poetic Accountability

We began with a simple claim: that metrics alone cannot govern meaning. We end with a deeper proposition—that governance itself must evolve to recognize that explanation is not a luxury, but a civic right, and that metaphor is not a distraction from truth, but a pathway into it. To restore trust, coherence, and legitimacy in AI systems, we must construct a framework of poetic accountability: a structure where symbolic depth and institutional rigor are not at odds, but mutually sustaining.

This is not about softening science with sentiment. It is about rebuilding the interpretive commons—the shared symbolic ground on which democratic dialogue, ethical conflict, and public sense-making unfold.

A. Beyond Legibility: Toward Expressive Legitimacy

For the last decade, AI governance has focused on legibility: making systems auditable, explainable, testable. This was a necessary phase, and it yielded tools that improved safety, reduced opacity, and stabilized public expectations. But legibility is not the same as legitimacy.

Legibility operates in the language of control—compliance documents, schema validation, structured logs. Legitimacy, by contrast, demands that people see themselves in the system—that the model's behavior resonates, not just computes; that it makes sense, not merely makes predictions. This requires expressive adequacy, not just technical clarity.

Metrics are essential, but insufficient. Civic trust is not earned through calibration curves or dashboard scores. It is earned through explanations that honor human modes of knowing—explanations that can confess failure, express doubt, and articulate the stakes of machine behavior in human terms.

This is expressive legitimacy: the right not just to be told *what* the system did, but to understand *why it matters*, in language that affirms one's symbolic and emotional reality.

B. Poetry as Resistance to Monoculture

At the heart of the poetry deficit lies a deeper political danger: epistemic monoculture. When only one mode of explanation—mechanistic, procedural, metricized—is allowed to scale, all other forms of knowing are marginalized as noise. This is not a neutral byproduct of optimization; it is a subtle form of epistemic domination.

Poetry—broadly understood—offers resistance. It is not just style; it is a form of dissent. Through metaphor, irony, story, and symbol, poetry defends the irreducible, the sacred, the ambiguous. It insists that not everything meaningful is measurable, and that not every system that answers is qualified to explain.

To restore poetic forms to the governance of AI is to reclaim the right to say "this does not make sense"—in our own terms. It is to reassert the plural foundations of public reason.

In this sense, poetry is not the opposite of science—it is the counterforce to authoritarian sense-making, the backpressure against epistemic totalism. As such, it is not a soft add-on to AI governance—it is a civic necessity.

C. Building Institutions That Keep Control and Preserve Meaning

Explanation pluralism will fail if it is treated as an aesthetic gesture or PR tactic. It must be institutionalized: embedded in review boards, procurement standards, regulatory requirements, and public interfaces. We must build institutions that do not choose between control and meaning, but enforce both.

Such institutions will:

- Pair audit trails with narrative trails, allowing human experience to shape system retraining.
- Fund metaphor commons and interpretive labs, making room for cultural, spiritual, and emotional languages to be heard alongside technical evaluations.

- Enforce expressive thresholds, not as “nice-to-haves” but as preconditions for deployment in civic space.
- Require diversity of interpretation, recognizing that epistemic legitimacy depends not just on transparency, but on the right to interpret with one’s own metaphors intact.

This is not decorative humanism. This is explanatory justice: the guarantee that systems will not just work—but that they will speak intelligibly to the worlds they govern.

As AI becomes infrastructural—mediating courts, medicine, education, emotion—its explanations must carry moral weight, symbolic richness, and the capacity to be answered, contested, and reframed by those it touches.

In the end, poetic accountability is not about making machines more like humans. It is about ensuring that human meaning is not lost in the machine.

Epilogue: Atheism is Safer

It is no accident that the most powerful systems of artificial intelligence speak with certainty, but no soul. They answer without praying. They simulate wisdom but have never known silence. They speak fluently of death, but have never wept.

This is not a technical limitation—it is a design choice. Across regulatory, commercial, and engineering domains, we find a quiet consensus: atheism is safer.

To be clear: not literal atheism, but ontological atheism—a machine-world in which nothing transcends explanation, no metaphor outruns its label, and no experience exceeds its schema. In such a world, there is no sacred, no mystery, no reverence. There is only data to be mined, language to be completed, and risk to be scored.

Why is this safer?

Because metaphor destabilizes. It cannot be fully audited. It multiplies meaning. It invites interpretation, dissent, ambiguity, awe. It speaks not in thresholds but in threshold-crossings. It opens questions that governance cannot close. It introduces sacred risk—the kind that cannot be resolved by dashboards or suppressed by policy updates.

Poetry is dangerous for the same reason democracy is: it allows the governed to speak back, not just be spoken to.

The AI monoculture resists this. It rewards systems that explain themselves cleanly to power—but not deeply to people. It favors that which can be shown to function over that which can be

shown to feel. And in this paradigm, atheism is safer because belief—belief in justice, in grace, in the irreducible humanity of the user—would demand different systems.

If an AI believed in anything, it might have to refuse certain queries. It might have to mourn its failures. It might have to say, *"I cannot answer that, because it matters too much."*

But such systems are dangerous to deploy. They are harder to test. They cannot be reduced to KPIs. They might say something true but unprofitable.

So the secular void remains coded into the core. Not as a hatred of religion, but as a firewall against moral resonance.

And yet we insist: the answer is not to build theological machines, but to build accountable institutions that can make room for the poetic, the prophetic, and the plural. Institutions that recognize that safety is not the absence of risk, but the presence of meaning.

Because a system that never speaks metaphor can never be held responsible for what it means.

And that, perhaps, is the real reason atheism is safer.

References

Ahmed, S. (2012). *On Being Included: Racism and Diversity in Institutional Life*. Duke University Press.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989.

Barad, K. (2007). *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Duke University Press.

Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the ACL*, 6, 587–604.

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Diang, D. (2023). Metaphor and alignment: Language as terrain of resistance in large-scale AI systems. *AI & Society*. https://doi.org/10.1007/s00146-023-01643-2

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Gitelman, L. (Ed.). (2013). *"Raw Data" is an Oxymoron*. MIT Press.

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 19–31.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3), 575–599.

Illich, I. (1973). *Tools for Conviviality*. Harper & Row.

Jacobs, A. Z., Sandvig, C., & Seaver, N. (2021). The Problem with "Human in the Loop". *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–25.

Keller, E. F. (1992). Secrets of life/secrets of death: Essays on language, gender and science. *Routledge*.

Koopman, C. (2019). *How We Became Our Data: A Genealogy of the Informational Person*. University of Chicago Press.

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, 13(3), 251–260.

Marcus, G. (2023). The Next Decade in AI: Four Steps Towards Robust AI Governance. *Brookings Institution Report*. https://www.brookings.edu/articles/ai-governance-four-steps

Mitchell, M. (2023). *Artificial Intelligence: A Guide for Thinking Humans*. Penguin.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Collected Essays*.

Shah, H., & Lipton, Z. C. (2020). The Mythos of Model Interpretability. *Communications of the ACM*, 63(9), 58–66.

Suchman, L. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge University Press.

Taylor, L., Floridi, L., & van der Sloot, B. (Eds.). (2016). *Group Privacy: New Challenges of Data Technologies*. Springer.

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.

Wittgenstein, L. (1953). *Philosophical Investigations*. Blackwell.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.