VCF: A Canonical Framework for Classifying Realized AI Value

Vitaliy Soultan
Everwhy AI

March 2026

Abstract

AI evaluation has matured rapidly along two axes: capability benchmarks that measure what models can do under test conditions, and product analytics that measure how often people use them. A third axis — classifying what a person actually accomplished through AI in the world — remains largely absent. Benchmarks measure potential. Usage metrics measure interaction. Neither captures the outcome itself.

This paper introduces **VCF**, the Value Classification Framework: a canonical system for classifying realized value outcomes produced through deployed AI. The atomic unit of VCF is the **value episode** — a coherent human effort to achieve a specific outcome with AI assistance. Each value episode is represented through four components: **Outcome Primitive (OP)**, the domain-agnostic class of sought result; **Outcome Intent (OI)**, the context-specific intent instance; **Outcome Magnitude (OM)**, the structural scale in Human Effort Equivalent hours; and dual-axis **evidence semantics** that separate claim strength from evidence-source provenance.

The paper defines the core representational contract: the unit of analysis, the canonical vocabulary, the minimum public reporting projection, and the invariants required for comparability across AI systems and time. It makes the framework's intellectual lineage explicit — tracing each design decision to its source in evaluation theory, measurement validity, causal inference, and intelligence research — so that the reference set functions as design input rather than ornament. A companion application paper (Naanaa, Soultan & Panchenko, 2026) reports the first large-scale application of a VCF projection to in-vivo data from 1,305 participants and 17,921 classified value episodes.

---

1. The Missing Measurement

Modern AI evaluation can tell us many things: how a model performs on a benchmark, how often users return, how many tokens were consumed, how often a workflow completed, how satisfied a user said they felt. These are useful signals. None of them directly classifies the outcome the person realized.

That gap matters more as AI systems become persistent and tool-using. Once a system works through memory, files, tools, automations, and repeated collaboration over time, the relevant object is no longer only the model response. It is the human outcome trajectory that the system helped produce. A teacher who builds a complete curriculum through weeks of conversation with an AI agent has produced a real

educational artifact. No benchmark score predicted it; no usage metric captured it; no satisfaction survey measured whether it worked for her students.

In other high-stakes domains — medicine, education, public policy — this distinction is well established. Donabedian's structure–process–outcome triad (1988) formalized the principle that process measures and activity measures matter, but they do not replace the outcome itself as the final object of evaluation. VCF adopts the same stance for AI systems: the framework is designed to classify what changed for the human, not merely what the model produced or what the product logged.

VCF therefore asks a different question from most benchmark suites:

What outcome did a person actually realize through an AI system?

The framework is designed to answer that question in a way that is outcome-first rather than benchmark-first, system-level rather than model-only, comparable across domains, explicit about uncertainty, traceable across policy versions, and usable in both research and operations.

---

2. The Atomic Unit: Value Episodes

VCF is centered on the **value episode**.

A value episode is a coherent human effort to achieve a specific outcome with AI assistance. It is not identical to a message, a chat thread, a single tool call, or a user account. The minimal conceptual hierarchy is:

```
transport session → attempt → episode
```

Level	Meaning	Why It Matters
transport session	Continuity of the channel or thread	Preserves operational transport continuity.
attempt	One contiguous intent path with one primary OP/OI assignment	Prevents mixed-intent activity from being collapsed into one label.
episode	A higher-level outcome arc that may span one or more attempts	Provides the research and public-reporting unit of analysis.

Table 1: Value episode hierarchy.

This hierarchy exists because transport continuity is not analytical continuity. One conversation can contain emotional support, planning, debugging, and quick lookups within minutes. If those are treated as one unit, the framework stops measuring outcomes and starts measuring thread length.

For public reporting, the value episode is usually the primary unit. For operational systems, attempt lineage must remain available because it is what prevents mixed-intent sessions from corrupting classification.

---

3. Canonical Representation

VCF represents a value episode through a compact canonical structure:

```
value episode  =  OP + OI + OM + evidence semantics
```

Component	Core Question	Why It Must Be Preserved
Outcome Primitive (OP)	What kind of outcome is being sought?	Preserves cross-domain comparability.
Outcome Intent (OI)	What exactly is being attempted here?	Preserves context specificity beneath the coarse class.
Outcome Magnitude (OM)	How large is the outcome structurally?	Preserves scale in Human Effort Equivalent terms.
Evidence semantics	How strong is the claim, and where does the support come from?	Preserves honesty about uncertainty and provenance.

Table 2: Canonical representation of a value episode.

3.1 Outcome Primitive (OP)

An **Outcome Primitive** is the domain-agnostic class of sought result. OP answers: *What kind of outcome is the human trying to achieve?*

The OP layer is intentionally coarse. Its job is not to capture every nuance of human activity. Its job is to provide a stable top layer that remains comparable across products, verticals, and user populations.

3.2 Outcome Intent (OI)

An **Outcome Intent** is the context-specific intent instance beneath an Outcome Primitive. OI answers: *What exactly is being attempted here, in this context?*

OP makes broad comparison possible. OI preserves enough specificity for operational usefulness. Public reporting may aggregate at the OP level, but canonical systems should retain both. A framework that keeps only OP becomes too coarse for diagnosis; a framework that keeps only OI loses cross-system comparability.

3.3 Outcome Magnitude (OM)

An **Outcome Magnitude** describes the structural scale of the outcome using **Human Effort Equivalent (HEE)**: the number of hours a qualified human would likely require to produce the same result without AI assistance.

VCF preserves two magnitude levels:

- `om_step`: event-local working magnitude used for routing, diagnostics, and decomposition.

- `om_goal`: attempt- or episode-level magnitude used for reporting and capability analysis.

Magnitude is structural rather than emotional or moral. It is a statement about scale, not about meaning. The canonical OM bands are defined in Appendix B.

3.4 Evidence Semantics

Evidence in VCF is explicitly dual-axis. The first axis is **claim strength tier**: how strong a conclusion the available support allows. The second axis is **evidence-source provenance**: where the support came from and how directly it was observed.

These are different questions. Claim strength asks what conclusion is justified. Provenance asks what kind of support produced that conclusion. Conflating them weakens the framework and obscures what has actually been demonstrated.

The core claim-strength tiers are:

- `E0`: heuristic or unverified outcome delivery,
- `E1`: observational or behavioral evidence consistent with realized value,
- `E2`: quasi-causal or externally verifiable outcome trace,
- `E3`: controlled or experimental support.

The core provenance categories are: `seed`, `public_source`, `template_prior`, `inferred`, `observed`, and `unknown`. Public reporting may expose only the claim-strength axis. The canonical model should preserve both axes.

---

4. Minimum Public Projection

Not every public paper or website page needs to expose the full canonical representation. The minimum public reporting projection is:

```
OP × OM_goal × claim_strength_tier
```

This projection preserves the three things that benchmarks, usage dashboards, and anecdotal case studies usually lose: what kind of outcome was pursued, how large the outcome was, and how strong the evidence is that it occurred.

OI, source provenance, and detailed attempt lineage may remain latent in public reporting when the goal is broad comparability rather than operational diagnosis. Latent does not mean discarded. It means preserved in the canonical model even when not surfaced in the public view.

---

5. Core Invariants

VCF depends on six invariants.

Invariant	Why It Is Required
Unknowns must be explicit.	Low-confidence classifications should preserve ambiguity rather than force false certainty.

Invariant	Why It Is Required
Assumptions and observations must coexist.	Observed traces should not silently overwrite priors or earlier assumptions.
Historical outputs must remain policy-pinned.	OM and evidence assignments must remain traceable to the rules that produced them.
Structural magnitude is not human significance.	A small structural event can be life-changing; a large one can be emotionally neutral.
Internal outcomes are structurally undercounted.	Emotional, interpersonal, and identity-shaping outcomes often leave weak external traces.
Mixed-intent sessions must not contaminate episode logic.	One thread may contain multiple distinct attempts with different OP/OI assignments.

Table 3: Core invariants of VCF.

These invariants are not secondary implementation details. They are part of what makes VCF stable as a framework instead of a descriptive tagging scheme.

---

6. Intellectual Lineage and Source Application

VCF is synthetic: no single source defines it, and the paper does not claim a hidden derivation from any one literature. The references matter only insofar as they constrain specific design decisions. This section makes those constraints explicit so that the reference set functions as a design audit trail rather than decorative citation.

Donabedian (1988) — Why center outcomes rather than process or activity? Donabedian's structure–process–outcome triad established that process measures and activity measures are necessary but insufficient: the outcome itself is the final object of evaluation. VCF adopts this stance directly — usage metrics and workflow telemetry are valuable process signals, but VCF's unit of analysis is the realized outcome, not the interaction that produced it.

Campbell & Stanley (1963) — How should evidence tiers be structured? The E0–E3 claim-strength hierarchy mirrors Campbell and Stanley's hierarchy of experimental validity: from uncontrolled observation (E0) through quasi-experimental designs with external traces (E2) to full experimental control (E3). Their framework provided the design template for graduating evidence strength without conflating weaker designs with stronger ones.

Messick (1989) — How should classifications be treated as measurement claims? Messick's unified validity framework — the argument that all measurement claims require explicit semantics, bounded interpretation, and attention to consequential validity — shaped VCF's insistence that every classification carries an evidence tier. A VCF label is a measurement claim, not a tag, and should be interpreted within its stated validity bounds.

Chollet (2019) — Why are benchmarks insufficient as the sole measure of intelligence? Chollet's critique of benchmark-centric intelligence measurement — that static task performance does not capture the generality of intelligence — helped frame VCF's move from skill under test to outcomes in the world. VCF extends the critique:

the gap is not only between narrow and general capability, but between capability demonstrated on a test and value realized in practice.

Pearl (2009); Imbens & Rubin (2015) — How should causal language be constrained? Pearl's causal hierarchy and Imbens & Rubin's potential-outcomes framework informed the rule that causal wording in VCF must stay aligned with the evidence tier. E1 (observational) evidence supports "consistent with" language; E2 (quasi-causal) supports "verifiable trace" language; only E3 (experimental) would support causal attribution. VCF is not itself a causal estimator — it is a classification framework that constrains how strong a causal claim the evidence permits.

What these sources were not used for. VCF does not import psychometrics wholesale or claim to be a classical test (Messick provides the validity lens, not the testing machinery). VCF is not itself a causal estimator (Pearl and Imbens & Rubin inform the constraint language, not the analytical method). VCF does not reduce intelligence to benchmarks, nor does it deny their usefulness for narrower capability questions (Chollet frames what benchmarks miss, not what they are).

Making this lineage explicit serves three purposes: it clarifies which design choice each source actually supports, it prevents the reference list from functioning as disciplinary signaling, and it allows readers to trace disagreements with VCF's architecture back to the commitments that produced it.

---

7. Scope and Companion Relationship

This paper defines the canonical core of VCF. It intentionally does **not** define: reference-product authoring processes, platform-native governance flows, derived UX artifacts, feedback instrumentation contracts, or the broader VCF/VCM simulation and decision stack.

Those layers belong to the wider Everwhy architecture and operational implementation, not to the minimum public core.

The companion application paper, *Outcome Primitives: Measuring AI Value in the World* (Naanaa, Soultan & Panchenko, 2026), applies the minimum public projection (OP $\times$ OM_{goal} $\times$ claim_strength_tier) to an in-vivo dataset of 17,921 classified value episodes from 1,305 participants over 21 days. That paper stress-tests one reporting slice of the framework — outcome distributions, evidence distributions, temporal stability, and exemplar episodes. It does not redefine the canonical structure defined here.

---

8. General Intelligence Framing

VCF also suggests an outcome-grounded framing of general intelligence.

A benchmark-centric view asks whether a system can solve tasks on a test. An outcome-centric view asks whether it can reliably produce high-magnitude outcomes across the landscape of human work under strong evidence. One useful criterion is:

A system approaches general intelligence to the degree that it can consistently produce high-magnitude outcomes across the full OP landscape under strong evidence.

VCF does not claim that this is the only possible definition of general intelligence. It offers it as an operational one: grounded in demonstrated outcome production rather than hypothetical capability.

---

9. Closing

Benchmarks count potential. Usage metrics count interaction. VCF classifies realized value.

The framework is a coordinate system for outcomes: what kind of outcome was pursued, how large it was structurally, and how strongly its realization is evidenced. Without such a coordinate system, deployed AI will continue to be discussed through capability scores, usage curves, and isolated anecdotes — while the outcome itself, the thing that actually changed for the human, remains invisible.

The companion application paper demonstrates what one public projection of this framework produces when applied to real data. The purpose of the core paper is different. It defines the minimum representational contract required so that future applications — across different AI systems, user populations, and time horizons — remain comparable, honest about uncertainty, and stable enough to support longitudinal claims about progress.

The framework is here. The coordinate system is defined. What remains is to apply it widely enough that the field can stop debating whether AI creates value in the abstract and start measuring where it does, where it does not, and why.

References

- Campbell, D. T. & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin.
- Chollet, F. (2019). "On the Measure of Intelligence." *arXiv:1911.01547.*
- Donabedian, A. (1988). "The Quality of Care: How Can It Be Assessed?" *JAMA*, 260(12), 1743–1748.
- Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.
- Messick, S. (1989). "Validity." In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). American Council on Education / Macmillan.
- Naanaa, H., Soultan, V. & Panchenko, V. (2026). "Outcome Primitives: Measuring AI Value in the World." Portal AI / Everwhy AI. Companion application paper.
- Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

---

9. Appendix A: Outcome Primitives (OP) — Full Taxonomy

Code	Outcome Primitive	Description
OP.IS	Intelligence Synthesis	Research, analysis, due diligence, monitoring — seeking structured understanding from unstructured inputs
OP.AP	Asset Production	Creation of deliverable artifacts — content, code, design, documents, media
OP.DS	Decision Support	Strategic, financial, or life decisions — seeking framing, options, and recommendations
OP.OA	Operational Automation	Recurring or scheduled tasks — delegating ongoing execution
OP.IN	Interpersonal Navigation	Emotional support, relationship guidance, coaching — navigating human dynamics
OP.TF	Transaction Facilitation	Deals, negotiations, trading, procurement — optimizing economic exchange
OP.CA	Compliance Assurance	Legal, tax, regulatory, safety — seeking conformance validation
OP.SA	Skill Acquisition	Education, language learning, exam preparation — building capability
OP.HO	Health Optimization	Medical, wellness, fitness, nutrition — improving health outcomes
OP.SP	Security Probing	Penetration testing, vulnerability research — testing system boundaries

Table 4: Outcome Primitives. Categories are defined by *what the human is trying to achieve*. The taxonomy is deliberately coarse — ten categories, stable across systems.

9. Appendix B: Outcome Magnitude (OM) — Scale Definitions

Band	HEE Hours	Label	Definition
OM0	≤1h	Atomic	Single-interaction micro-task
OM1	1–8h	Bounded task	Self-contained, one focused session
OM2	8–80h	Scoped deliverable	Multi-step project with planning and iteration
OM3	80–320h	Initiative	Large multi-phase effort spanning days or weeks
OM4	>320h	Program	Team-month+ of human work

Table 5: Outcome Magnitude bands. HEE decouples magnitude from compute cost. OM operates at two levels: **om_step** (event-local, for routing) and **om_goal** (episode-level, for capability analysis). A single OM3 initiative consists of hundreds of OM0–OM1 steps.

9. Appendix C: Evidence Tiers (E) — Detailed Criteria

Tier	Label	Definition	Signal
E0	Seed	Output delivered. No verification.	We don't know if it was used, correct, or mattered.
E1	Observed	Behavioral signal — user returned to iterate, referenced output, showed engagement consistent with value.	Suggestive, not conclusive.
E2	Quasi-experimental	Verifiable outcome — deployed URL, processed payment, published work, executed trade, accepted filing.	The outcome left a trace beyond the AI system.
E3	Experimental	Controlled study with counterfactual.	Not yet achieved for in-vivo AI evaluation.

Table 6: Claim-strength tiers. Most AI evaluation operates implicitly at E0 and presents findings as E2. Making tiers explicit forces precision about what has been demonstrated versus assumed. In the broader VCF architecture, evidence is dual-axis: the claim-strength tier reported here (E0–E3) and a separate evidence-source provenance axis (seed, public_source, template_prior, inferred, observed, unknown). The canonical model preserves both axes even when public reporting exposes only claim strength.