AI as a Forcing Function for Shared Meaning:

Decision Traces and the Unacknowledged Decision Surface

Ron Itelman

Draft v3 — January 2026

Abstract

Decision intelligence offers rigorous methods for making decisions under uncertainty. Yet, in day-to-day practice with AI systems, a quieter problem emerges: people and systems act without sharing the same meaning — and nobody notices.

Warren Powell's unified framework for stochastic optimization is an exemplar of rigor in decision science. In principle, his formulation allows ambiguity about meaning to reside in the state. In practice, however, ambiguity about interpretation is seldom modeled explicitly. It lives in undocumented assumptions, user intuition, and silent defaults.

This work does not modify Powell's framework. Instead, it argues that decision science benefits when shared understanding becomes explicit and traceable. Decision traces operationalize a dimension of the state that is usually implicit: how meaning is negotiated, clarified, confirmed, and reused across time and systems.

The goal is to instrument ambiguity explicitly. By exposing ambiguity, tracing clarifications, and quantifying what uncertainty remains, we can reason about how understanding emerges — and identify the next most valuable question to ask. Decision traces make ambiguity observable, measurable, and governable, creating conditions for mutual symmetry of intent between human and system.

This paper does not assume semantic invariance a priori. Meaning is treated as something negotiated through interaction. Once that negotiation is recorded and shareable, downstream systems can reconstruct the reasoning and perform consistent transformations. In this sense, decision traces function as instrumentation that enables gauge-like stability across otherwise heterogeneous systems.

The contribution of this paper is conceptual and infrastructural: It proposes decision traces as a framework that standardizes how ambiguity in meaning is surfaced and documented before optimization, making the otherwise unacknowledged decision surface explicit.

1 Introduction

Organizations increasingly rely on AI systems to assist, automate, and justify decisions. Much of the discussion about risk focuses on data quality, model performance, fairness, or cybersecurity. These are real issues. But a quieter and more pervasive risk appears earlier in the decision process:

the human and the system do not share the same meaning — and nobody notices.

Consider something simple: a column labeled "temperature." In one system it means Celsius. In another, Fahrenheit. In another, it is undocumented.

The AI still answers the question:

"What is the average temperature?"

Dashboards update. Reports are produced. Decisions follow — all with confidence — even though the decision surface being optimized over was never agreed upon.

This problem sits upstream of model accuracy and model risk. It is not about whether the algorithm is correct, but whether the underlying interpretation is shared.

Two strands of work make this especially important.

First, Powell's unified framework for stochastic optimization reminds us that every decision depends on the definition of the state. If the state is missing information, poorly specified, or semantically ambiguous, every downstream optimization inherits that distortion — often invisibly.

Second, emerging work on "System 0" highlights that AI systems increasingly function as part of our cognitive infrastructure. We begin to trust them by default. As trust increases, we ask fewer clarifying questions about what the system actually meant.

The result is a new category of operational risk:

- systems act before meaning is shared,
- teams assume alignment that does not exist,
- trust grows faster than understanding.

Traditional responses — better prompts, more dashboards, tighter controls — do not address the core issue. What is missing is instrumentation.

We need systems that:

- hold ambiguity instead of prematurely collapsing it,
- treat unclear fields as unknown rather than "close enough,"
- ask diagnostic questions to surface hidden assumptions,
- establish shared intent before committing to action.

In short, trust in AI should not mean believing the answer. **Trust should mean shared meaning before action.**

This paper proposes *decision traces* as a way to make that process observable, measurable, and governable.

2 What this Paper Does and Does Not Claim

This work does not modify Powell's mathematics or propose a new decision-theoretic model. His framework already presumes that the relevant state and data are correctly specified.

The contribution of this paper concerns how these frameworks are applied in operational settings. In practice, there is rarely an explicit step that requires practitioners to surface, disambiguate, and document the meanings they are assuming when they specify the state.

As a result, ambiguity in interpretation remains invisible, even though optimization proceeds as if it were resolved. This paper describes decision traces as a mechanism to standardize this missing step.

Decision traces make the interpretive layer explicit, checkable, and reusable — turning meaning into something that can be governed rather than silently assumed. In this sense, the contribution is infrastructural: extending how decision intelligence is practiced, not the mathematics beneath it.

Decision traces and the notion of the unacknowledged decision surface introduced in this paper provide a way to formalize the instrumentation and processes through which meaning is clarified, recorded, and reused across systems that feed decision models.

Related Work and Novel Contributions

What is novel in this paper is the claim that, in AI-enhanced "System 0" settings, ambiguity about meaning should be treated as an explicit part of the decision state rather than left implicit. When this ambiguity is not made visible, systems operate over what this paper calls *unacknowledged decision surfaces*: hidden assumptions about interpretation that propagate through downstream optimization.

Using the simple example of three tables and a user query to an AI agent, we show how decision traces provide a practical way to surface and record this semantic uncertainty before action is taken.

A Minimal Example: One Term, Two Possible Meanings

Consider a column labeled "temperature." In practice, it may refer to the same physical quantity expressed in different units:

$$U_{\text{temp}} \in \{\text{Celsius},\ \text{Fahrenheit}\}.$$

Before clarification, the system does not know which interpretation is correct. This is represented as a belief distribution:

$$p(U_{\text{temp}}) = \begin{bmatrix} p(\text{Celsius}) \\ p(\text{Fahrenheit}) \end{bmatrix}.$$

If we had a justified reason to treat the two possibilities as equally likely, we might write, for illustrative purposes:

$$p(U_{\text{temp}}) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}.$$

It is important, however, not to confuse this notation with knowledge. In many practical settings we do not in fact know that the prior should be uniform — nor do we know whether it should be 0.9/0.1, 0.7/0.3, or something else entirely. The assumption of a particular prior is itself a decision.

In such cases, the appropriate representation is not a probability distribution but simply a set of candidate meanings. Decision traces make this explicit: they record whether probabilities were assigned, on what basis, and at what point in the process, rather than smuggling a default prior into the model as a silent assumption.

A clarifying question is then asked:

> "Are these values recorded in Celsius or Fahrenheit?"

Suppose the user confirms Celsius. We update the belief:

$$p(U_{\text{temp}} \mid \text{answer}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Uncertainty has not disappeared magically; it has been resolved because the human and the system now share the same meaning.

Three Tables, One Question

We now connect the simple example to a more realistic scenario, closer to the diagram motivating this work.

Suppose we have three datasets that all contain a column labeled "temperature":

- Dataset 1: a CSV uploaded by the user (no documented units).
- Dataset 2: an internal system table where "temperature" is known to be Celsius.
- Dataset 3: an internal system table where "temperature" is known to be Fahrenheit.

For each dataset $d \in \{1, 2, 3\}$, we define a latent semantic variable

$$U_d \in \{\text{Celsius}, \text{Fahrenheit}\}.$$

For the system-owned tables, the meaning is fixed:

$$p(U_2) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{(Celsius)}, \qquad p(U_3) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{(Fahrenheit)}.$$

For the user-uploaded CSV, the meaning is unknown. We may illustrate this with a uniform prior merely for exposition,

$$p(U_1) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix},$$

but in practice there may be no justified basis for assigning these weights. What matters is not that the system chooses a particular prior, but that any such choice is made explicit, traceable, and revisable. When no basis exists, the ambiguity should be represented simply as an unresolved set of candidate meanings rather than as a numeric belief.

The user now asks:

> "What is the average temperature?"

At this moment, there are *two layers of uncertainty*:

1. **Semantic uncertainty** about U_1 (is the user's CSV in Celsius or Fahrenheit?).
2. **Interpretation uncertainty** about the agent's behavior: will it check units, reconcile semantics across tables, or silently assume one?

If the agent simply averages values across datasets without resolving U_1, the reported "average temperature" rests on an unacknowledged decision surface: the system has acted as if a particular meaning were true, without ever negotiating that meaning with the user.

Legend.

- d — dataset index (1: user CSV, 2: system Celsius, 3: system Fahrenheit).
- U_d — latent semantic meaning of "temperature" in dataset d.
- $p(U_d)$ — belief about the meaning in dataset d.
- $[1, 0]$ — certainly Celsius; $[0, 1]$ — certainly Fahrenheit.
- **Semantic uncertainty** — not knowing U_1.
- **Interpretation uncertainty** — not knowing how the agent treats U_1 when answering.

3 System 0: Meaning, Trust, and Cognitive Infrastructure

Chiriatti et al. argue that data-driven AI systems now function as a distinct layer of cognition, which they call *System 0*. Unlike Systems 1 and 2, which

describe human reasoning, System 0 reflects the outsourcing of perception, retrieval, filtering, and recommendation to machine systems that operate continuously in the background.

As AI becomes cognitive infrastructure, users increasingly trust its outputs by default. This reduces incentives to clarify meaning, particularly in routine operational contexts — precisely where hidden semantic assumptions can propagate the farthest.

4 Meaning Provenance

Modern data systems track where data comes from: lineage graphs, ETL pipelines, and version histories. What they rarely track is how meaning is established. When fields are ambiguous, contexts shift, or assumptions go unstated, the interpretation layer disappears — even though it determines every decision that follows.

Meaning provenance records how meaning is proposed, questioned, clarified, confirmed, and reused. Decision traces capture this process. Rather than treating interpretation as invisible, it is treated as a first-class infrastructure.

5 Implications to the Field of Decision Theory

This paper has argued that a significant class of decision risk arises not from poor models, weak data, or malicious actors, but from something more basic: the absence of shared meaning at the moment of decision.

Using a minimal example — a column labeled "temperature" — this example shows how ambiguity in semantic interpretation can be represented as a belief distribution and reduced through clarifying questions. This framing emphasizes that ambiguity is not an error to be ignored, but a state of knowledge to be instrumented.

The three-table scenario illustrates how such ambiguity appears in practice. A user uploads a dataset with unknown units, while internal systems contain data labeled in Celsius and Fahrenheit. When an AI system answers the question "What is the average temperature?", it may silently assume a meaning without ever negotiating it. The result is an unacknowledged decision surface: the system optimizes over an implicit interpretation the user may not share.

Within Powell's unified framework for stochastic optimization, this form of uncertainty belongs in the state. In principle, the ambiguity over semantic meaning — and even uncertainty about the agent's interpretation policy — can be modeled and reasoned about. In practice, however, these uncertainties are almost never made explicit. They persist as undocumented assumptions, hidden defaults, and quietly inherited risks.

At the same time, the emergence of "System 0" reminds us that AI is no longer a tool sitting beside human cognition. It is becoming part of our cognitive infrastructure. As trust increases, clarification decreases. This amplifies the

cost of ambiguity precisely at the point where organizations most depend on automation.

Decision traces offer one path forward. By recording how meaning is proposed, questioned, clarified, confirmed, and reused, they introduce *meaning provenance* as first-class infrastructure. Ambiguity can be exposed, quantified, and governed. The system can identify the next most informative question before committing to an answer. Shared meaning precedes action.

The ideas in this paper do not claim to replace existing decision frameworks. Rather, they extend them. By making semantic uncertainty visible, traceable, and measurable, decision traces provide the instrumentation required to align human intent, system interpretation, and downstream optimization.

In short, if decisions are made on top of meaning, then meaning itself must become something that can be observed, reason about, and designed for.

6 Decision Traces Capture Reasoning Step Data

To make this idea concrete, Table 1 illustrates the kinds of artifacts that a decision trace captures. The goal is not to model cognition directly, but to record the interpretive steps that shape the decision.

Element	**Example**
What was ambiguous	`temperature: {C, F, unknown}`
What candidates existed	`{unit=C, unit=F}`
What questions were asked	"Are these Celsius?"
Who answered	user / system / default
What changed	Fahrenheit removed; Celsius confirmed
What assumptions were chosen	"Assume Celsius unless corrected"
What uncertainty remained	unresolved fields; conflicting context
Where meaning was reused	downstream pipeline; later query

Table 1: Illustrative components of a decision trace.

Decision traces do not attempt to model cognition directly. They are not an ontology and they are not an effort to reconstruct mental processes from logs of human–AI dialogue. Instead, they make latent interpretive work visible in real time. The formalism provides shared steps through which the user and the system articulate, question, confirm, and revise the meanings that shape a decision.

A key distinction is perspective: decision traces are generated during interaction, across populations of conversations, and can be diffed, compared, and reused. The structure emerges from the bottom up, as ambiguity is surfaced, negotiated, and resolved. In this way, decision traces create reusable data about

reasoning without requiring access to a system's internal chain of thought, while remaining compatible with downstream statistical analysis when appropriate.

7 What is Meant by a Decision Surface

In decision science, every decision is made with respect to a model of the world. Informally, we can think of the *decision surface* as the space the system believes it is optimizing over.

This notion appears in many forms:

- the boundary of a classifier,
- the landscape of a value function,
- the mapping from states to actions induced by a policy.

In Powell's notation, a policy π selects an action on the basis of the current state S_t:

$$a_t = \pi(S_t).$$

All quantities of interest — decisions, rewards, transitions — are defined over this state. Consequently, the decision surface depends entirely on what is represented inside S_t.

When Meaning Is Missing From the State

If part of the relevant world is missing from the state, the system is still capable of optimizing — but it is optimizing in the wrong space. This observation is not controversial; it is foundational to decision theory.

The argument advanced in this paper is simply that *semantic ambiguity belongs to this category.* Meaning is part of the real decision problem, yet in most operational systems it is treated as metadata, external to the state, and therefore invisible to the policy.

Unacknowledged Decision Surfaces

This paper uses the term *unacknowledged decision surface* to describe the situation in which:

1. ambiguity about meaning exists,
2. the system silently commits to one interpretation,
3. neither the system nor the user recognizes this commitment, and
4. optimization proceeds as though the interpretation were certain.

The point is not that the decision surface is unknown. Rather, it is *implicitly chosen* without ever being interrogated or disclosed.

In the formalism introduced earlier, let U denote an unknown semantic variable with belief distribution $p(U)$. If the system acts without clarification, it effectively collapses to some assumed value u^* and then optimizes conditioned on that assumption. This is exactly what we mean by "unacknowledged."

What Is New Here

The claims made here are not that existing decision frameworks are incorrect, nor that new mathematics is required. What is new is the emphasis on three ideas:

1. **Framing.** Semantic meaning should be treated as part of the state, not as auxiliary metadata.
2. **Instrumentation.** Once ambiguity is in the state, it can be measured, traced, and governed through decision traces.
3. **Language.** Naming these situations as *unacknowledged decision surfaces* makes their risk visible and discussable.

In this sense, the contribution is conceptual rather than algebraic: this paper extends the scope of what is considered "in the state," and provides practical tools to make that extension operational.

This work sits at the intersection of decision theory, data governance, and human–AI interaction, connecting ideas that are often treated separately.

References

[1] Powell, W. B. (2019). A unified framework for stochastic optimization. *European Journal of Operational Research*, 275(3), 795–821. doi:10.1016/j.ejor.2018.07.014

[2] Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as system 0 thinking. *Nature Human Behaviour*, 8, 1829–1830.