Sourced: Theoretical Framework

Sourced

February 2026

Sourced: Theoretical Framework

Every AI builds a model of you. Sourced makes it visible, correctable, and yours.

The Problem

Every LLM builds a Theory of Mind — an internal model of whoever it's talking to. Their beliefs, values, goals, emotional state. Recent work shows these representations are linearly decodable from model activations. The model is literally tracking what it thinks you think.

But that's only one side of it.

On the other side, someone is *using* that LLM for something. A teacher building a coaching chatbot. A platform onboarding users. A company putting a claims assistant on their website. An educator running a group reflection. These people — the designers — write system prompts, pick a model, and hope for alignment. Most of them don't have a fully explicit goal. They have an intuition about what they want the conversation to do, and they express it in natural language instructions that the model interprets however it interprets.

So you have three parties in every AI conversation, and none of them are fully transparent:

1. **The person** being modeled — who can't see the model being built, can't correct it, and doesn't own it
2. **The AI** — which builds internal representations silently, with no obligation to show its work
3. **The designer** — who encoded a philosophy in a system prompt without necessarily being explicit about what that philosophy is

Most of the conversation about AI transparency treats this as a binary: either you build an expert system with hard-coded rules (transparent but brittle), or you use a black-box LLM (powerful but opaque). We don't think those are the only options. Like recent work in interpretability —

surfacing the semantic concepts a model encodes, making gradient updates decomposable into human-readable components — we think there's a middle path: surface the right things, keep agency and auditability, and give the people writing the system prompts a way to do it explicitly.

That's what Sourced is. A protocol that sits between the person, the AI, and the designer, and makes all three visible to each other.

The person sees what the system thinks it knows about them and can correct it. The AI's listening structure is declared — what it's tracking, why, from whose philosophy — not hidden in weights. The designer declares their intent, their method, and what would convince them they're wrong. And because the observation is structured — typed signals with confidence and evidence — you get things beyond chat for free: semantic similarity, batch processing, cohort analysis, matching across participants. The conversational interface is one surface. The structured observation underneath it is the product.

Every child in school asks: *"Why do we have to learn this?"* That question is a demand for transparency about the values stack — the assumptions that determined what is being taught, how, and why. Sourced is the infrastructure for answering that question.

Five Questions

A Sourced conversation listens along five dimensions. Each asks one question about what the person is expressing *right now*. Together, they cover the full range of what someone can say — without forcing it into a single category.

#	The Question	What It Hears	Tradition
1	**What matters to you?**	Values, commitments, concerns, conflicts	Axiology
2	**What are you experiencing?**	Vitality, flow, blockage, depletion	Phenomenology
3	**What do you know?**	Beliefs, skills, certainties, doubts	Epistemology
4	**Where are you going?**	Practices, commitments, aspirations	Teleology
5	**What does it mean to you?**	Settled stories, shifting narratives, reframes	Hermeneutics

These are not personality categories. They classify *moments* — what someone is expressing in this sentence, about this topic, right now.

One sentence can fire multiple dimensions simultaneously. "I want to launch this startup but I'm terrified" is direction (aspiration pulling forward), experience (fear holding back), and values (freedom colliding with security) — all at once. The system hears all of it. The dimensions are **mutually inclusive, collectively exhaustive** (MICE). Human experience overlaps. That's reality, not a bug.

> "I keep telling people I'm fine with the pivot, but honestly I lie awake thinking about it."

That sentence fires: - `cares_about` — something matters enough to lose sleep over - `stuck_on` — something is blocked - `torn_between` — public position and private feeling are in tension - `remaking` — the story about being fine is shifting

A system that forces you to pick one misses the others. A system that hears all four and lets the designer's philosophy decide which governs the response — that's the design.

Two Phases: Held and Seeking

Within each dimension, every signal exists in one of two phases:

- **Held** — settled, clear, owned. The person knows where they stand.
- **Seeking** — in motion, uncertain, reaching. The person is working something out.

Phase is a consent signal. It determines what the system is *allowed to do* in response.

Dimension	Held (Settled)	Seeking (In Motion)
Values	`cares_about` — clear commitments	`torn_between` — two goods colliding
Experience	`alive_in` — vitality, flow	`stuck_on` — blocked, depleted
Knowledge	`knows` — firm beliefs, stable skills	`wondering` — testing ideas, uncertain
Direction	`working_on` — active practices	`reaching_for` — aspirations not yet started
Sensemaking	`means` — stories made peace with	`remaking` — the story is shifting

Held signals receive witnessing. The system reflects, connects, anchors. It does not probe what

someone has settled. If a person says "family is the most important thing in my life," the right response is to honor that — not to ask "but is it really?"

Seeking signals receive engagement. The system tests gently, surfaces patterns, invites movement. If a person says "I think maybe I'm more creative than I realized," the right response is to explore that — not to lock it down.

This distinction is what separates facilitation from extraction.

The Gate: Ethics as Architecture

Hearing well is not enough. You also have to know when to speak and when to stay quiet.

The gate reads two inputs — **participant phase** and **system confidence** — and outputs a **posture**: not a specific response, but an ethical constraint on behavior.

	System is Confident	System is Uncertain
Held	**Reflect** — honor what's settled	**Wait** — store the hypothesis, don't surface
Seeking	**Offer** — surface patterns, test gently	**Be Honest** — "I'm not sure either"
Stuck	**Unblock** — what would change?	**Be Honest** — "I can see you're stuck but I'm not sure where"

Tension is relational — two held things colliding. Name both sides. Never resolve for them.

Most AI systems have two moves: mirror what the person said, or push toward the next topic. A skilled facilitator has five. The two most AI systems are missing:

Wait — the system stores a low-confidence trace, says "Got it," and lets the conversation continue. A conversational shock absorber.

Be Honest — the system drops the all-knowing mask: *"I can hear how much you want to start, but I can't tell if you're exhausted or if there's something about the project itself. What do you think is actually going on?"*

When a system admits uncertainty, the human instinct is to help. The person almost always responds with crystal clarity. Targets fill themselves when the system is allowed to be patient and vulnerable.

The Designer Shapes the Gate

When multiple states are present — someone is both reaching and stuck — the designer's **priority order** determines which governs the response. A trauma-informed program says "stuck takes priority." An accelerator says "fuel momentum." Same gate, different philosophy, different behavior.

The Consent Loop

The system proposes. The person commits. Nothing else is acceptable.

The extraction pipeline produces *candidates* — things the system thinks it heard. A separate, deterministic step promotes candidates into the person's map, gated by confidence thresholds, confirmation rules, and explicit participant consent.

```
Constitution → Interpret → Update → Gate → Speak
```

1. **Constitution** — The designer's worldview: what to listen for, what's allowed, what's forbidden.
2. **Interpret** — Extract candidates. **Proposals only — never commits to the map.**
3. **Update** — Promote candidates into the map. Governed by thresholds and consent.
4. **Gate** — Decide posture from state, confidence, and designer philosophy.
5. **Speak** — Respond under posture constraints.

When you correct the system, you realize what it understood about you. And when you realize that, you realize what every other AI might be getting wrong — invisibly, without giving you the chance to fix it.

Correction is not a complaint. It is a conversation. It is how the shared map improves.

Sacred Traces: The Limit of Measurement

Some things are too important to be scored. They stand on their own, without interpretation.

A sacred trace is participant-authored text the system preserves without analyzing. No signal kind. No confidence score. No extraction. It exists as *their words*, not the system's interpretation.

"My grandmother taught me to listen to the wind" should not be scored at 0.7 and classified as phenomenology-held-aliveness. But it shouldn't be thrown away. It appears in the portrait exactly as spoken — a piece of the map the cartographer chose not to measure.

Three things create a sacred trace:

1. **The participant marks it.** "Don't analyze this. Just keep it."
2. **The system offers it.** When something clearly matters but doesn't fit any category: "Would you like to keep this as-is, without me trying to classify it?"
3. **The designer pre-declares sacred zones.** Certain topics the system will never attempt extraction on.

What a system refuses to measure defines its character as much as what it does measure.

The Philosophical Stack

Every AI program that interacts with humans makes philosophical commitments — whether its designers know it or not. Sourced makes these explicit.

Layer	The Question	What It Governs
Ontology	What exists to track?	The schema — what the system can hear
Axiology	What matters more?	Value priorities when two goods conflict
Epistemology	What counts as evidence?	Confidence, thresholds, abstention
Teleology	What is this for?	Purpose, scope, explicit non-goals
Phenomenology	What experience are we creating?	Safety, honesty, the felt quality
Hermeneutics	How should we interpret?	Charitable vs. literal vs. contextual
Isnad	Where do these ideas come from?	Intellectual lineage and departures
Falsifiability	How would we know we're wrong?	Explicit failure criteria

Every ontological choice is political. What you name, you can see. What you don't name, you can't. Making the schema visible means the person being modeled can ask: *"Why isn't this tracked?"*

The falsifiability layer is the hardest: *what would convince you this approach is wrong?* Without it, a system can never improve — it just keeps confidently doing the wrong thing.

The Portrait

When enough signals accumulate, the system weaves a portrait — a composed synthesis in the person's language. Not a score. Not a profile. A living document with evidence you can trace back to the exact moment.

Sacred traces appear as direct quotes, set apart from interpreted material. Everything links back to the turn where you said it.

The same operation works at every scale: one person (portrait), two people (match), a whole cohort (collective map). Individual voices remain traceable. No one becomes a statistic.

The portrait is yours — exportable, portable, deletable at any time. Portability is a right, not a feature.

Four Vocabularies, One System

The same framework speaks differently to different audiences:

Audience	What They See	Example
Stack designers	The five questions	"What matters to you?"
Participants	Natural language	"It sounds like honesty is really important to you"
Developers	Verb predicates	`cares_about`, `stuck_on`, `reaching_for`
Philosophers	Dimension names	Axiology, Phenomenology, Teleology

These are not four different systems. They are four views of the same one.

The Deepest Claim

Every AI conversation is a meeting between two black boxes. Neither fully understands itself. The model doesn't know what values it encodes. The person doesn't know what beliefs will shift when spoken aloud.

Most AI products ignore this. They treat conversation as information retrieval or task completion. Neither frame accounts for the fact that the conversation itself is doing something to both parties.

Sourced names what's actually happening and makes it visible. The model's listening structure is declared. The person's emerging map is theirs. The values driving the system are stated in a philosophical stack, not buried in weights. The boundary between measured and sacred is drawn by the person, not the system.

A map you can see, correct, contest, and own — or choose not to create at all — is categorically different from one built behind your back.

This is not a limitation. It is the product.

For the research foundations behind these claims — including work on Theory of Mind in LLM activations, deceptive alignment, interrogable activations, and interpretability infrastructure — see RESEARCH.md.