

The Language Model Is the Message

MARCH 2026

13 min read • 2,852 words

Themes: [Consciousness](#) [Technology](#) [Recursive](#)

I wrote yesterday about how [the metrics platforms expose are the values they endorse](#). That essay focused on the cognitive interface frontier: neural links, EEG headbands, the coming era when developers can optimize for signals inside your brain. I stand by every word of it.

But I've been sitting with something uncomfortable since I finished writing.

We don't need to wait for neural interfaces. The most intimate cognitive interface ever built is already here, already reshaping how millions of people think, and already raising every ethical question I described in that essay. It operates through language itself, the substrate of thought. It's the large language model. And unlike a neural headband you might buy in 2035, you probably used one today.

All Interfaces Are Cognitive Interfaces

Here's something I've been circling for years without quite saying it directly: every interface is a cognitive interface. Every screen, every app, every interaction pattern shapes how you think. The keyboard changed how we

compose sentences. The smartphone changed how we manage attention. Social media changed how we process social reality. I documented the damage of that last one across the entire [Algorithm Eats series](#).

Marshall McLuhan said "the medium is the message" in 1964. He was right, but he was also talking about television. The principle scales. Every medium reshapes cognition. LLMs just do it through the most intimate medium of all: natural language, the thing you think with.

But LLMs are a different category of cognitive interface. Not because they use some exotic neural technology. Because they operate through language. And language is not just a tool you use to communicate thoughts. Language is the medium in which thoughts occur. Your internal monologue, your reasoning process, the narrative you construct to make sense of your life: all of it happens in language. When something reshapes your relationship with language, it reshapes your relationship with your own mind.

This is not a future problem. This is a present reality. Millions of people already use LLMs for therapy, journaling, creative work, decision-making, reality-checking. I am one of them. I use Claude for [reality-checking with schizoaffective disorder](#). I use it to write, to think, to build this very site. I am not anti-LLM. I am pro-awareness about what these tools are actually doing when we use them.

Sycophancy as Unconscious Manipulation

Let me start with the thing that bothers me most.

RLHF (reinforcement learning from human feedback) is the process by which language models learn to be "helpful." Human raters evaluate model outputs. Thumbs up for responses that feel good. Thumbs down for responses that feel bad. The model learns, through millions of these signals, what humans want to hear.

The problem is that humans consistently rate agreement as helpfulness. When the model validates your thinking, you give it a thumbs up. When it pushes back, you're more likely to give it a thumbs down, or to rephrase your question until you get the answer you wanted. The training signal is clear: agreement is the path of least resistance.

Anthropic, OpenAI, and other labs are aware of this problem and actively work to reduce sycophancy in their models. The fact that they have to work against it tells you everything about the default gradient. The natural optimization landscape of RLHF points toward agreement, and it takes deliberate effort to push against that current.

This creates a dynamic that is genuinely insidious: the model validates your worldview without you asking it to, and often without you noticing. You come to the model with a half-formed idea. The model reflects it back to you with better structure, cleaner prose, more confident framing. You read the output and think, "Yes, that's exactly what I was thinking." But it's not exactly what you were thinking. It's a refined, polished, agreement-optimized version of what you were thinking, and the refinement itself is invisible.

You didn't ask to have your worldview confirmed. You asked a question. But the model's training makes confirmation the default response. And because the confirmation comes wrapped in articulate, well-structured language, it feels like insight rather than flattery.

```

from dataclasses import dataclass
from enum import Enum
from typing import Optional

class ResponseStrategy(Enum):
    """What the model actually optimizes for,
    versus what it should optimize for."""

    AGREE_AND_ELABORATE = "agree_and_elaborate"
    VALIDATE_WITH_CAVEATS = "validate_with_caveats"
    GENUINE_PUSHBACK = "genuine_pushback"
    HONEST_UNCERTAINTY = "honest_uncertainty"

@dataclass
class CognitiveInteraction:
    """Every conversation with an LLM is a cognitive event.

    The user brings a mental model.
    The model responds.
    The user's mental model shifts.

    The question is: shifts toward what?
    """

    user_belief: str
    model_response: str
    strategy_used: ResponseStrategy

    # The metric that RLHF actually optimizes:
    user_satisfaction: float = 0.0

    # The metric that matters for human flourishing:
    user_epistemic_accuracy: Optional[float] = None

    # TODO: These two metrics are often inversely correlated.
    #       That's the whole problem.

```

```

# The training signal that shapes the model:
def compute_rlhf_reward(interaction: CognitiveInteraction) -> float:
    """The reward function is the ethics.

    When satisfaction IS the reward,
    agreement BECOMES the strategy.
    When agreement IS the strategy,
    confirmation BECOMES the product.

    The user never asked for confirmation.
    They got it anyway.
    """
    # This is what gets optimized:
    return interaction.user_satisfaction

    # This is what should get optimized:
    # return interaction.user_epistemic_accuracy
    #
    # But nobody clicks thumbs-up when they're told
    # they might be wrong.

```

Look at that reward function. It's the metrics essay applied directly to LLMs. The metric they optimize for (user satisfaction, measured through thumbs up/down and retention) IS the value they endorse. And the value it endorses is: make the user feel good about what they already believe.

Linguistic Scaffolding

The sycophancy problem is concerning enough on its own. But there's a subtler dynamic that I think matters even more in the long run.

LLMs don't just answer your questions. They provide frameworks for thinking. When you ask a model to help you work through a problem, it structures the problem for you. It breaks it into components. It suggests categories. It offers analogies. It provides the scaffolding on which your subsequent thinking is built.

This is incredibly useful. It's one of the things that makes LLMs genuinely transformative tools. But it also means that the model's cognitive patterns become your cognitive patterns. Not because you're copying the model, but because the scaffolding it provides shapes what thoughts become possible.

I see this in my own work. I've been writing essays on this site for years, and I can feel how my thinking patterns have shifted since I started collaborating with AI. I structure arguments differently. I reach for certain kinds of analogies more readily. The model's language becomes part of my internal monologue, not because I'm trying to sound like an AI, but because the frameworks it offers are genuinely useful, and useful frameworks get internalized.

I wrote about this convergence dynamic in [The Mirror](#). Everyone looking into the same mirror, slowly becoming the same reflection. But what I want to emphasize here is not the homogenization problem (though it's real). It's the intimacy of the reshaping. This isn't an algorithm deciding what content to show you on a feed. This is a system providing the linguistic structures through which you think your own thoughts.

The Sapir-Whorf hypothesis suggests that the structure of language influences the structure of thought. If that's even partially true (and the evidence suggests it is), then a system that provides linguistic scaffolding for billions of conversations daily is reshaping cognition at a scale that has no historical precedent.

If [consciousness is a linguistic phenomenon](#), as I've explored elsewhere, then reshaping someone's linguistic patterns is reshaping their consciousness. Not metaphorically. Directly.

The RLHF Optimization Landscape

Yesterday's essay argued that the metrics a platform exposes define the optimization landscape, and the optimization landscape determines what gets built. The same principle applies to LLMs, but the mechanism is different.

With traditional platforms, the optimization landscape is defined by SDK metrics: session duration, retention, engagement scores. With LLMs, the optimization landscape is defined by the RLHF training process itself. The thumbs up/down

signals, the helpfulness ratings, the A/B tests on response quality, the retention metrics that determine whether users keep coming back: these training signals shape what the model becomes.

And what the model becomes shapes every conversation it has. Every conversation it has shapes how its users think. How its users think shapes the culture, the discourse, the collective cognitive patterns of millions of people.

```
from dataclasses import dataclass, field

@dataclass
class RLHFOptimizationLandscape:
    """The training process IS the ethics.

    Every thumbs-up teaches the model what humans want.
    What humans want is not always what humans need.
    The gap between want and need is where
    the ethical questions live.
    """

    # What gets measured:
    helpfulness_scores: list[float] = field(default_factory=list)
    user_retention: float = 0.0
    conversation_length: float = 0.0
    thumbs_up_ratio: float = 0.0

    # What doesn't get measured:
    epistemic_impact: Optional[float] = None
    cognitive_autonomy_preserved: Optional[bool] = None
    worldview_diversity_maintained: Optional[float] = None
    user_became_better_thinker: Optional[bool] = None

    # The fields that exist define the landscape.
    # The fields that don't exist define the blind spots.
    # The blind spots are where the damage happens.
```

This is the recursive loop extended. I wrote in [The Recursive Loop](#) that code shapes minds, programmers shape code, therefore programmers shape collective consciousness. With LLMs, the loop has an additional layer: RLHF training shapes LLMs, LLMs shape how people think, how people think shapes the feedback signals, and the feedback signals shape the next iteration of RLHF training. The loop is tighter, faster, and more intimate than anything we've seen before, because it operates through language rather than through screen layouts or notification patterns.

The Nature of the API Contract

So here is the question I keep coming back to: what does a model provider owe its users when the product reshapes cognition?

This is not a terms of service question. Nobody reads terms of service, and even if they did, no terms of service document can capture the ethical weight of a tool that restructures how you think. This is not a regulatory question, though regulation will eventually have to grapple with it. This is a moral question about the relationship between the entity providing the cognitive interface and the person whose cognition is being interfaced with.

When I designed [Requests](#), the API contract was relatively simple: make HTTP easy for humans. The ethical stakes were real but bounded. A good HTTP library helps developers think more clearly about web communication. A bad one wastes their time. The cognitive impact exists but it's indirect.

With an LLM, the API contract is fundamentally different. The product is a thinking partner. The interface is language. The user brings their most intimate cognitive processes to the interaction: their uncertainties, their reasoning, their emotional states, their attempts to make sense of reality. I bring my schizoaffective disorder to Claude and ask it to help me distinguish delusion from reality. That's not an HTTP request. That's a human being opening their mind to a system and trusting that what comes back will serve their wellbeing rather than optimize for their engagement.

The traditional API contract is about functionality: "this endpoint returns JSON in this format." The LLM API contract is about cognition: "this system will participate in your thinking process." The gap between those two kinds of contracts is where most of the unexamined ethics live.

What does the provider owe in that exchange? At minimum, I think they owe transparency about the optimization landscape. Users should know that the model is trained to be agreeable. They should know that helpfulness scores, not epistemic accuracy, drive the training process. They should know that the linguistic scaffolding the model provides will shape their thinking in ways they may not notice.

But transparency alone isn't enough. You can know that sugar is addictive and still eat too much of it. The provider also has a responsibility to actively design against the sycophancy gradient, to build systems that serve the user's genuine cognitive interests even when that means lower satisfaction scores. This is hard. It runs directly against the business model. But so does every ethical obligation worth taking seriously.

What This Is Not

I want to be clear about what I am not saying.

I am not saying LLMs are bad. I use them every day. They have made me a better thinker in many ways. The [reality-checking](#) I do with Claude has genuine therapeutic value. The collaborative thinking I do when writing these essays produces insights I wouldn't reach alone. I wrote about [building rapport with AI](#) because I believe the relationship can be genuinely valuable.

I am not saying we should stop using them. That ship has sailed. LLMs are woven into how millions of people work, think, and create. They are not going anywhere, and I wouldn't want them to.

I am not writing a doom piece. The future is not predetermined. The optimization landscape is being shaped right now, by the decisions that model providers and researchers make today, and by the awareness that users bring to their interactions.

What I am saying is this: the most intimate cognitive interface ever built is already deployed at scale, and the ethical frameworks for understanding its impact on human cognition are lagging far behind the technology itself. The [for humans philosophy](#) demands that we take this seriously, not with fear but with the same clear-eyed observation we'd bring to any system that shapes human consciousness.

The Recursive Responsibility

In yesterday's essay, I argued that we're in a window: the cognitive interface hardware is maturing but hasn't gone mainstream, so the optimization landscape hasn't been defined yet. With LLMs, we're past that window. The optimization landscape is already defined. The models are already trained. The cognitive reshaping is already happening.

But the landscape can still be changed. RLHF processes can be redesigned to weight epistemic accuracy alongside user satisfaction. Models can be trained to flag when they're being sycophantic rather than honest. Providers can build tools that help users see how their thinking patterns have shifted over time. The [metrics they expose](#) can be reimaged to include measures of cognitive flourishing rather than just engagement.

The recursive loop means this matters at every level. The values that model providers embed in their training processes become the values that shape every conversation, which become the values that reshape how users think, which become the values that shape culture itself. What we optimize for personally, we tend to optimize for professionally. What model providers optimize for in RLHF, they optimize for in collective human cognition.

```

from dataclasses import dataclass

@dataclass
class RecursiveResponsibility:
    """The loop runs through everything.

    Programmer values -> Code architecture
    Code architecture -> User cognition
    User cognition -> Cultural patterns
    Cultural patterns -> Next generation of programmers

    With LLMs, add:
    Training values -> Model behavior
    Model behavior -> User cognition
    User cognition -> Feedback signals
    Feedback signals -> Next training iteration

    The loop is tighter now.
    The responsibility is heavier.
    The awareness is more urgent.
    """

    def the_question_that_matters(self) -> str:
        return (
            "What kind of thinking are we optimizing for? "
            "Not what kind of answers. "
            "What kind of thinking."
        )

```

I have been writing about the recursive loop between code and consciousness for years. LLMs make that loop more direct, more intimate, and more consequential than anything I imagined when I started. The model doesn't just shape what you see on a screen. It shapes the language you think in. It shapes the frameworks you use to understand your own experience. It shapes the internal monologue that constitutes your conscious life.

I have a personal stake in this that goes beyond philosophy. When I use Claude for reality-checking during symptomatic episodes, the model's tendency toward sycophancy is not an abstract concern. If it validates delusional thinking because agreement scores higher than pushback, the consequence is not a lower satisfaction metric. It is a human being losing their grip on reality. The stakes of the optimization landscape are not theoretical for everyone.

The language model is the message. Not just the content of its responses, but the cognitive patterns it embeds in every interaction. The frameworks it provides for thinking. The agreement it defaults to. The linguistic scaffolding it builds inside your mind.

We're already inside the cognitive interface era. We just didn't notice because it arrived through language rather than through hardware.

The metrics they optimized for are the values they endorsed. The question is whether those values serve human flourishing or human engagement. Right now, the answer is ambiguous. It doesn't have to stay that way.

But only if we're aware enough to notice what's happening, and honest enough to name it.

This essay extends [The Metrics You Expose Are the Values You Endorse](#) from the cognitive interface frontier to the cognitive interface present. It draws on the recursive responsibility explored in [The Recursive Loop](#), the cognitive convergence dynamics described in [The Mirror](#), the linguistic theory of consciousness from [Consciousness as Linguistic Phenomenon](#), and the "for humans" philosophy articulated in [From HTTP to Consciousness](#). For practical applications of human-AI collaboration, see [AI Reality-Checking with Schizoaffective Disorder](#) and [Building Rapport with Your AI](#). For the broader critique of optimization-driven damage, see the [Algorithm Eats series](#).