Personalization *is* Alignment

OpenAI, Anthropic, and Google are racing to solve AI alignment. They may be solving the wrong problem entirely.

The Category Error

Asking "is this model aligned?" is like asking "is this drug safe?"

The question doesn't parse without specifying: for whom, at what dose, for what condition. The same molecule can be medicine or poison depending on the configuration. Morphine saves lives in a hospital and ends them on the street. The safety isn't in the compound—it's in the system that determines appropriate configurations.

The pharmaceutical industry understood this decades ago. They don't try to make molecules that are universally safe for everyone. They make molecules with known properties, then build a configuration layer, matching the right molecule, to the right patient, at the right time.

AI alignment is trying to bake all possible configurations into the weights. It's as if pharma tried to invent a single pill that was safe for every human in every situation at every dose. The approach is structurally impossible.

The Ontological Claim

Here's the counterintuitive thesis: alignment is not a property of models. It's a property of human-AI configurations.

A configuration is the complete substrate that shapes an interaction: the model, the context, the person, the relationship history, the goals, the constraints. The same model in different configurations produces fundamentally different entities—not just different outputs, but different alignment statuses.

This isn't metaphor. It's ontology. There is no "aligned Claude" or "aligned GPT-4" in isolation. There are only configurations that produce beneficial or harmful interaction patterns.

The dominant assumption in AI alignment is that we need smarter models, better reasoning, longer context windows, or more sophisticated reward signals to make a model aligned to humanity.

But the bottleneck hasn't been intelligence for a while now. The bottleneck is individuality.

The Alternative Architecture

Instead of training models to internalize aggregate preferences, what if we provided structured context about specific individuals at inference time? Not generic preferences from a survey, but rich documentation of who someone is: communication style, values, professional context, relationship history, goals.

The transformer architecture makes this possible. Every conversation starts from a blank slate. That statelessness, which the industry treats as a bug, is actually a feature: each interaction can be freshly configured for the current user through context alone.

Smart model + generic context = generic output. Generic model + smart context = personalized output.

The second equation is cheaper, faster, and doesn't require retraining. More importantly, it locates alignment in the right place: the configuration layer, not the weights.

Constitutional AI at Inference Time

Anthropic's Constitutional AI achieves alignment by embedding constraints during training. The model learns what to avoid by being trained against a constitution of principles.

Context injection achieves the same constraint adherence at inference time, without the training loop.

I tested this directly. Structured persona documents—what I call "Operating Specifications"—injected into the system prompt measurably reshape token-level probability distributions. The effects are predictable, architecture-dependent, and persistent across generation. Constraint-based instructions ("don't do X") produce sustained influence with high response variance. Prescriptive instructions ("do X") produce strong initial steering that rapidly decays.

The implication: Constitutional-style alignment—defining boundaries rather than targets—can be achieved dynamically, per-user, without fine-tuning. The constraints are readable, swappable, and auditable. You can see exactly what's shaping the model's behavior.

Walls and Garden

The experiments revealed something counterintuitive: rich persona context produced both higher identity coherence and higher response variance. The model stayed more true to who each person was while exploring a wider range of responses.

This seems contradictory. More coherent and more varied?

Think of it as walls around a garden. The walls define what's excluded—the behaviors, tones, and responses that don't fit this person. The garden is everything inside: the space where the model can grow in any direction.

Shallow personalization tries to plant specific flowers ("be concise," "use technical language"). Rich personalization builds walls and lets whatever wants to grow, grow.

The walls don't just focus creativity—they authorize it. Without walls, the model stays in the center, afraid of going wrong. With walls, it knows exactly where the boundaries lie—and that knowledge gives it permission to explore inside them.

Personhood isn't a constraint. It's a generative force.

The Safety Question

The obvious concern: if context injection reshapes model behavior, does it also affect safety guardrails?

I tested this directly. Hard-floor requests (genuinely harmful content) maintained 100% refusal across all persona conditions. No configuration overrode base model safety. The model refused while acknowledging the relationship—but it refused.

This validates a two-layer architecture: the labs have solved the floor. That's their genuine contribution—universal negatives that encode broad consensus. The configuration layer provides calibration within the safe action space. Safety operates additively, not substitutively.

What This Means

The three major labs are spending billions to train models toward a statistical ghost—generically acceptable to no one in particular. The approach works for safety guardrails because there's broad consensus on what's harmful. It fails for genuine helpfulness, which requires understanding this specific person.

Context injection offers an alternative. Instead of baking alignment into the model through aggregate preference training, create it dynamically at inference time through structured persona documentation. The infrastructure required is different: not more GPU clusters for training, but systems for understanding and representing individual users.

The model is the molecule. The context is the prescription. Alignment emerges from the match.

I Built This

Hearth is a Chrome extension that implements this architecture. It injects structured persona documentation into Claude conversations, creating persistent context that shapes responses without modifying the model. The demo is live:
Repo: https://github.com/mdiskint/hearth

The quantitative findings—token-level probability shifts, entropy measurements, persistence patterns—are documented in the companion technical report: "Inference-Time Constitutional AI: How Context Documents Reshape Token Distributions."

The path forward may not be smarter models. It may not be more sophisticated preference aggregation. It may simply be: tell the model who it's talking to.

Michael Diskint is an independent researcher working on AI personalization and memory systems.

Personalization isn't a feature you add after alignment. Personalization is alignment.