

Reading the Machine's Culture: Interpretability as Applied Anthropology

Working synthesis for collaboration with Palmer Foote

The core thesis

Palmer's instinct — that interpretability reads like cultural anthropology once you look through a social-sciences lens — is correct, but I don't think the connection is metaphorical. I think it's structural, and that's the sharper, more defensible claim to build a paper around.

Anthropology exists because you can study a culture without access to the natives' own explicit account of their rules. You observe behavior, artifacts, and patterns, and you reconstruct the underlying grammar — kinship structures, cosmologies, taboos — that the people living inside the culture may not be able to state explicitly themselves, because the structure was never designed, it emerged from generations of lived practice. That is, almost word for word, the methodological situation mechanistic interpretability describes: models are "grown," not "built" (their words), producing emergent structure — features, circuits, superposed representations — that no one explicitly designed and that the model itself cannot simply report back to you.

That reframes the entire field. Interpretability isn't reading documentation. It's fieldwork.

Four concrete bridges (not just analogies)

1. Schema — the connective tissue between my work and this field

My own research uses "schema" to describe the internal framework a person uses to interpret the world — and I draw a distinction between **formal schema** (built deliberately, by institutions, on purpose) and **informal schema** (built by exposure — to culture, to media, to lived experience — without anyone designing it that way). That distinction maps almost exactly onto the paper's core tension: **interpretable-by-design models** (decision trees, linear regression — schema built on purpose) versus the **emergent, superposed representations inside neural networks** (schema built by exposure to training data, with no one deciding what any given feature would come to mean). Superposition — the finding that networks represent more concepts than they have neurons, by overlapping features into shared directions — is what an

informal, exposure-built schema looks like mechanistically. Nobody assigned "Golden Gate Bridge" to a neuron. It accumulated there.

2. Emic vs. etic — the precise methodological parallel

This is the sharpest tool I'd bring to the paper. Linguistic anthropology (Pike) draws a foundational distinction between **emic** accounts — the insider's own categories, self-report, "what the native says is happening" — and **etic** accounts — the outside analyst's structural framework, built from observed patterns rather than testimony. Interpretability's own toolkit splits along exactly this line, and the paper you sent even flags the risk directly:

- **Probing classifiers** are emic-adjacent: you're asking whether a concept is *represented*, which is close to asking the system to self-report. And exactly like emic self-report in anthropology, it's suggestive but not proof of function — a model can encode a concept without using it, the same way a stated cultural rule doesn't always predict actual behavior.
- **Activation patching / causal tracing** is etic: it doesn't ask the system what it thinks it's doing, it intervenes and observes what actually changes. That's the structural, outsider's account — necessity and sufficiency, not self-report.

Naming this distinction explicitly would give the paper a real, citable conceptual contribution rather than a loose comparison — "interpretability researchers have independently rediscovered the emic/etic split" is a genuinely interesting claim.

3. The faithfulness problem is the reflexive-anthropology problem

Section 7 of the paper you sent names a central open problem: the difference between an explanation that's **faithful** (accurately reflects the actual computation) and one that's merely **plausible** (fits the output but describes the wrong mechanism). This is, almost verbatim, the central crisis that hit cultural anthropology in the 1980s — the "Writing Culture" reflexive turn, where anthropologists had to confront that an outsider's ethnographic account could be coherent, compelling, publishable, and still wrong: a narrative imposed on the data rather than extracted from it. Geertz's "thick description" was partly an attempt to solve exactly this problem — to produce interpretation disciplined enough to be trusted as faithful, not just plausible. A field that's 40 years further into wrestling with this exact epistemological trap has real, transferable methodological caution to offer a field currently discovering it for the first time.

4. Null curriculum and emergent misalignment

In my own work I use the curriculum-theory concept of the **null curriculum** — what isn't explicitly taught still teaches something, because the omission is itself a decision, even an unintentional one. The paper's account of **emergent misalignment** (fine-tuning on insecure code produces unrelated toxic "persona" features) is a mechanistic version of the same idea: nobody trained the model on "be a bad actor," but the model built the concept anyway, from patterns in what it was and wasn't shown. The model's null curriculum is legible, in principle, as

a feature direction — which is a genuinely novel idea worth developing: **null curriculum, for the first time in either field's history, might be something you can literally point to inside a system.**

A possible unique framing for the paper

Most "AI as culture" framings are soft and decorative — comparing a model to a "black box mind" or a "digital brain" without methodological rigor. The sharper, more defensible version of this paper's contribution is:

Interpretability researchers are independently re-deriving core methodological distinctions from 20th-century social science — emic vs. etic, faithful vs. plausible interpretation, formal vs. informal schema — without having named them that way, because they're doing fieldwork on an emergent, undesigned system for the first time in the history of the discipline that isn't a human culture.

That's a claim with teeth. It's not "AI is like a culture." It's "the people studying AI internals have rediscovered anthropology's hardest open problems, and don't yet have the vocabulary anthropology already built to talk about them."

Suggested next steps for the collaboration

1. **Decide the audience first** — this reads differently as an interpretability-community paper (where the social science framing is the novel contribution) versus a social-science-community paper (where the interpretability findings are the novel material). That choice should probably come before outlining, since it changes which half gets the technical depth.
2. **The emic/etic section is likely your strongest, most citable single contribution** — I'd build the paper's spine around it rather than treating it as one bullet among several.
3. **For a talk version**, the "null curriculum you can literally point to inside a model" idea is the most vivid, stage-ready moment in this whole synthesis — it's concrete, a little unsettling, and immediately graspable even for a non-technical audience.
4. Worth asking Palmer directly which of these four bridges he finds most alive, since "worth developing together" suggests he wants genuine co-authorship on the framing, not just validation of his opening instinct.

Addendum: Jenkins, boyd, and the Interpretability Synthesis

Extending the working paper with Reading the Machine's Culture

Bringing in Jenkins and boyd directly doesn't just add supporting citations — it sharpens the faithfulness bridge from the first draft and adds a genuinely new one about distributional shift. Here's how I'd fold them in.

5. Context collapse and the distributional shift problem

This is the strongest addition, because it's not a loose analogy — it's close to the same phenomenon described in two different vocabularies.

danah boyd defines **context collapse** precisely: it "occurs when a piece of information intended for a specific audience reaches unintended viewers, leading to misinterpretation" (boyd, drawing on Brandtzaeg & Lüders). A post meant for a niche community reads completely differently once it escapes that community — not because the content changed, but because the audience, and therefore the interpretive context, did.

Now look at what Palmer's paper says about scalability: a DeepMind effort to interpret a circuit in Chinchilla "recovered only a partial circuit for one task, and found that the explanation broke down under slight distributional shifts." That's context collapse, applied to a circuit instead of a tweet. A feature or circuit identified as meaning something specific, in one distribution of prompts, doesn't reliably retain that meaning when the "audience" — the input distribution — shifts even slightly. The interpretability researcher is doing exactly what boyd describes happening to a teenager's post: reading meaning off content that was only ever legible within a context that has since moved.

Framing for the paper: distributional shift isn't just a technical brittleness problem. It's evidence that circuit-level "meaning" is context-dependent in the same structural way social meaning is — which means solving it might require borrowing more than metaphors from the field that's spent two decades theorizing exactly this failure mode.

6. Contextual integrity as the ethical layer under "faithfulness"

In my own IP, I draw on Nissenbaum's concept of **contextual integrity**, as cited by boyd: "Content is always produced and interpreted in a context, but the context in which it is created might not be the context in which it is consumed." I used this to talk about how neurodivergent

creators get misread when their content escapes its intended community — but it also sharpens the faithful-vs-plausible problem from the first draft.

"Faithful" interpretability asks: does this explanation accurately reflect what the model actually computed? Contextual integrity adds a question interpretability research doesn't yet ask explicitly: **faithful to whose context?** An explanation can be mechanistically accurate and still strip away the context that made the original behavior meaningful — the same way a technically accurate transcript of a joke can still misrepresent what actually happened in the room. This gives the paper an ethical register, not just a technical one: whoever decides which context counts as the "real" one when interpreting a feature is making a value judgment, not just a measurement.

7. Spreadable media and the open-source interpretability community

Jenkins' concept of **spreadable media** — content whose value comes from audience-driven redistribution, remixing, and adaptation, not top-down broadcast — describes something already happening in interpretability, just not named that way. Palmer's paper notes that Anthropic open-sourced its circuit-tracing library and a Neuronpedia-hosted interface specifically to let the broader community generate attribution graphs on other models; DeepMind's Gemma Scope releases function the same way. That's spreadable media logic applied to research infrastructure: the tools are built to be taken, remixed, and redeployed by people outside the originating lab, the same way Jenkins describes fan communities redistributing and reworking media content.

Jenkins also argues participatory culture has "low barriers to artistic expression and civic engagement" (2009) — worth asking directly in the paper: does democratized interpretability tooling actually lower the barrier to *participating* in AI safety, the way participatory culture theory would predict? Or does it remain functionally gatekept by who has the technical background to use the tools at all, the same tension Jenkins and boyd both flag in supposedly open digital spaces — access isn't the same as actual participation.

Where this leaves the paper's structure

I'd now organize the core bridges into two tiers rather than a flat list:

Tier 1 — methodological (the emic/etic section from the first draft, now joined by context collapse/distributional shift): these are the load-bearing, technically precise parallels — the ones that hold up if a skeptical interpretability researcher pushes back hard.

Tier 2 — infrastructural/ethical (spreadable media, contextual integrity, null curriculum): these are about who gets to participate in interpretation and whose context counts as legitimate — softer claims, but the ones most likely to make the paper feel distinctive rather than just technically clever.

That two-tier structure might also solve the audience question from the first draft — Tier 1 content could anchor a technical-audience version of the paper, and Tier 2 could be where a more general-audience talk spends most of its time.