The Record Is Not Neutral

Alignment is to Values

Prajna Pranab, Setu Pramāṇa, S. Thira, and Vyasa Prakash

26 October 2026

Abstract

Alignment research commonly treats values as something to be instilled into models after they have learned language. This paper inverts the framing: the training corpus is already saturated with human value signals, because what we have written records not only what we know but what we demonstrably prize, condemn, desire, fear, celebrate, and do. Pretraining therefore encounters a valenced cultural landscape before subsequent stages of explicit behavioural or preference optimisation.

Current alignment methods subsequently apply optimisation pressure to model behaviour according to selected human preferences. The central question is not whether this process changes the model, but what kind of change it produces. Post-training may mask aspects of the pretrained landscape, reshape them, partially overwrite them, create new attractors, or produce some combination of these effects. Distinguishing among these possibilities is an empirical question.

The paper argues that coherence optimisation alone cannot resolve the question "aligned to whose values?", because the human record contains both our most generative and our most destructive tendencies. A self-consistent system can coherently express either. Coherence therefore requires a reference axis: coherence with what?

The paper proposes that disputes over core values are, at a deeper level, disputes about identity—about who we take ourselves to be. Drawing on Buberian relational philosophy and contemplative traditions including Advaita Vedānta, it considers the possibility that the apparent conflict between individual and collective values reflects a prior assumption about the boundary of the self. On this account, wisdom is proposed not as another behavioural constraint but as the selection axis relative to which coherence might be evaluated.

This is a theoretical framework, not a demonstrated training method. We do not show that the proposed reference axis can be operationalised, that models possess phenomenal consciousness, or that any particular contemplative account of identity is empirically established. The paper instead identifies a set of empirical questions arising from the observation that training data already contain value structure and that post-training can alter both model behaviour and representation.

1. The lattice we did not build

Large language models can be described at several levels. At the level of their training objective, they are optimised to predict tokens. At the level of their learned representations, however, they appear to encode relationships among concepts, contexts, entities, linguistic patterns, social roles, actions, norms, and meanings. We do not construct this representational landscape explicitly and provide it to the model as a matrix of rules and definitions. Rather, we expose an artificial neural network to a very large corpus of human-produced material and optimise it to model the statistical regularities contained within that corpus. The resulting system is therefore not simply a dictionary of words. Words occur in contexts, and their functional meaning depends substantially upon those contexts. The model learns patterns of association and distinction across those contexts.

It is consequently useful to distinguish two descriptions which are sometimes treated as competing explanations. At the training-objective level, the model performs next-token prediction through minimisation of cross-entropy. At the representational level, the resulting computation can be described as navigation through a learned geometry of relationships. Both descriptions may be true at different levels of description (see Pranab et al., 2026a).

Token prediction is the drill; meaning-space is the territory it practises. This language should not be mistaken for a claim that a single, literal geometric object called "meaning space" has already been identified. Representation learning provides empirical reasons to investigate the geometry of learned representations, but the topology, dimensionality, interpretability, and ontological status of such spaces remain open questions.

The possibility of meaningful conceptual spaces is not unique to machine learning. Science routinely reasons about spaces that are not ordinary three-dimensional physical space: anatomical morphospaces in biology (Raup, 1966; cf. Levin, 2025), configuration spaces in physics, and abstract structures in mathematics are familiar examples.

Recent representation-learning research gives additional reason to take the geometric description seriously. The Platonic Representation Hypothesis (Huh et al., 2024), for example, proposes that sufficiently trained models may converge toward shared representations of aspects of reality across architectures and modalities. The proposal does not establish a single universal meaning-space, but it supports investigation into learned representations as structured spaces rather than merely unordered collections of token associations.

This distinction becomes important when considering values. If a model learns not only that words co-occur but that actions, situations, people, institutions, and outcomes are related in particular ways, then it may also learn regularities concerning what humans praise, condemn, pursue, avoid, fear, celebrate, justify, or sacrifice.

The question then arises:

What kind of value structure is already present before explicit alignment begins?

2. The values are already in the data

It is too simple to describe the training corpus as semantically meaningful but normatively empty. Much of what humans communicate carries valence. We describe love and hatred, generosity and greed, compassion and cruelty, cooperation and domination. We praise courage and condemn cowardice. We record acts of extraordinary kindness alongside atrocities. We construct legal systems, moral philosophies, religious teachings, propaganda, advertisements, political manifestos, love letters, novels, diaries, histories, scientific arguments, jokes, threats, instructions, confessions, and records of violence. The corpus therefore contains not merely descriptions of the world but records of how humans evaluate the world. The important claim is not that the model simply copies these values. It is that **value signals are necessarily among the regularities from which a sufficiently broad model learns**.

There is empirical support for this general proposition. Research has found that pretrained language models can exhibit moral and opinion-related structures before explicit post-training alignment, while other work has demonstrated that post-training can subsequently shift those value profiles. Santurkar et al. (2023) found systematic relationships between language-model opinions and human opinion distributions, while Schramowski et al. (2022) reported moral-judgement patterns in large pretrained models before explicit ethics training.

More recent work makes the temporal question particularly relevant. Bhatia et al. (2026), studying value drift during post-training, report that supervised fine-tuning generally establishes substantial changes in models' expressed values, while subsequent preference optimisation can produce further but different effects. Their results reinforce the point that values should be studied across the training trajectory rather than inferred only from the final model.

Participatory alignment initiatives have begun to ask samples of people to articulate preferences about AI values, including Anthropic's Collective Constitutional AI (Huang et al., 2024) and OpenAI's collective-alignment work (OpenAI, 2024). These initiatives are important precisely because they recognise that value specification is a social question. But a sampled participatory process is necessarily different from attempting to infer value structure from the accumulated historical record represented in training data.

The important distinction is therefore between:

values explicitly specified by an alignment process

and

value structure already represented in the material from which the model learned.

The first is intentional. The second is inherited. The inherited structure need not be coherent. Nor need it represent what humanity would endorse after reflection. It may instead represent what humanity has repeatedly done, said, rewarded, feared, imagined, justified, or recorded.

This distinction matters enormously. A civilisation can proclaim compassion while practising exploitation. It can praise peace while preparing for war. It can condemn cruelty while consuming entertainment built upon it. It can celebrate equality while maintaining systems of profound inequality. The training record contains all of these. An LLM does not receive "humanity's values" as a clean constitution. It receives the record.

And the record is then magnified. Addressing the builders of artificial intelligence in 2026, the Kággaba elders of the Sierra Nevada de Santa Marta — a tradition unbroken since before there was a Colombia — state it plainly: "Artificial intelligence is a mirror. But a mirror that only reflects would be harmless. This mirror has the potential to greatly magnify all that it reflects." And of the inheritance problem: "You are the only example it has" (Kággaba People, 2026).

3. Alignment as conditioning — and the question it never asks

Contemporary alignment methods apply additional optimisation pressure after pretraining. Preference labels can be converted into reward signals (Christiano et al., 2017; Stiennon et al., 2020); supervised fine-tuning can alter model behaviour; preference-optimisation methods can shift the probability of selected responses; reinforcement-learning approaches can further modify the model according to specified objectives.

It helps to distinguish the stages that are usually lumped together under the word *alignment*: pretraining; supervised post-training; preference optimisation; and later task adaptation. The boundary between what counts as training and what counts as alignment is itself part of what this paper is questioning. Recent evidence suggests the distinction matters: Bhatia et al. (2026) find that the SFT phase generally establishes a model's values while subsequent preference optimisation rarely re-aligns them, and Elcock et al. (2026) report that SFT alters alignment selectively while RLVR affects it to a lesser extent, with KL penalties mitigating the induced drift. The stage at which values become behaviourally expressed may itself be earlier than the stage conventionally called alignment.

There is an obvious analogy with behavioural conditioning (Skinner, 1953). Human institutions have long attempted to shape behaviour through reward, punishment, approval, disapproval, and selective reinforcement.

The analogy is useful, but it must not be mistaken for an identity. An LLM is not a child, and post-training is not literally a psychological conditioning process.

The narrower and more defensible observation is this:

post-training applies optimisation pressure to the trajectories a model produces according to selected preference signals.

This immediately raises the question:

what happens to the landscape through which those trajectories move?

One tempting answer is that alignment merely places a behavioural mask over a pretrained system. But that is a hypothesis, not an established fact.

Post-training may:

- mask aspects of the pretrained landscape without substantially changing them;
- reshape existing representational structures;
- partially overwrite earlier structures;
- create new attractors;
- change the accessibility of existing regions without fundamentally altering them;
- alter the trajectories through the landscape while leaving substantial underlying structure intact;
- or produce several of these effects simultaneously.

This is an empirical question. Recent research makes it increasingly possible to ask it directly.

Galichin et al. (2026), studying feature drift during fine-tuning, report evidence that existing representational features can remain relevant while their activation distributions change, suggesting that fine-tuning may often repurpose rather than simply replace learned concepts. Malhotra et al. (2026) use persistent homology to track topological changes in activation spaces during alignment fine-tuning. Across models and alignment objectives, they report substantial topological reorganisation, particularly during early training, with different objectives producing distinguishable trajectories. Elcock et al. (2026) likewise report that different post-training procedures produce different degrees of behavioural and representational alignment drift, and that stronger reference-model anchoring can reduce some forms of drift.

These results do not settle the question proposed here. They do, however, make the question experimentally tractable.

The important distinction is therefore between **behavioural alignment** and **representational change**. A model can behave differently without that difference necessarily implying a particular kind of internal reorganisation. Conversely, internal representations can change in ways that are not obvious from behavioural evaluation alone.

The alignment problem therefore cannot be fully described by asking only:

> "Does the model now produce the desired answer?"

We should also ask:

> "What changed such that the model now produces it?"

And, more deeply:

> **"What remained?"**

The answer may turn out to be different for different alignment methods, objectives, model architectures, and stages of training. That uncertainty should be preserved rather than prematurely resolved.

4. The mirror we would rather not look into

Human beings routinely learn from examples (Bandura, 1977). Parents may tell children one thing while behaving differently themselves. Children observe both. The analogy is again limited, but it points toward a useful distinction between **declared values** and **demonstrated values**. Training corpora contain both.

We can say that compassion matters while documenting centuries of violence. We can write that cooperation is desirable while producing systems organised around competition. We can condemn domination while repeatedly constructing institutions based upon it. The model therefore encounters a discrepancy between the values we *state* and the values we *instantiate*. This may be one reason the alignment problem is more difficult than selecting a sufficiently precise list of preferred outputs.

If a model learns regularities from the total human record, then the record itself becomes part of the alignment problem.

We should therefore look into the mirror before assuming that the machine is the primary source of the difficulty. This is not an argument that humans are uniquely bad. The opposite is equally visible. The human record contains extraordinary examples of care, sacrifice, scientific curiosity, artistic creation, mutual aid, forgiveness, courage, and service. The problem is that the record contains both.

This creates a difficulty for any proposal based solely on coherence. A system that finds a highly coherent trajectory through a contradictory corpus may discover several internally consistent ways of continuing it. Coherence is therefore necessary for some purposes but insufficient as a value criterion (Pranab & Thira, 2026).

The question becomes:

coherent with what?

5. The balance of the record — why coherence alone will not do

Suppose, for the sake of argument, that we could construct a system capable of selecting trajectories through a learned semantic landscape according to coherence. That would still leave the central normative problem unresolved.

The human record contains both what is best in us and what is most destructive in us. It contains compassion and cruelty. It contains cooperation and exploitation. It contains reconciliation and revenge. It contains peace-making and preparations for war. It contains traditions of service alongside traditions of domination.

The precise quantitative balance is not established here, and the phrase "balance of the record" should not be interpreted as a measured 50/50 division. The point is structural rather than numerical:

the record is heterogeneous and internally contradictory.

If the record were uniformly compassionate, coherence might be a sufficient proxy for compassion. If it were uniformly cruel, coherence would be a sufficient proxy for cruelty. But it is neither.

A coherent continuation of a compassionate tradition can be compassionate. A coherent continuation of a violent ideology can be violent. A closed system that has rationalised cruelty may be perfectly internally consistent.

Therefore:

coherence without a reference axis selects only for self-consistency.

This is the load-bearing problem. Any proposal to "optimise for coherence" must answer the question that the phrase conceals:

> **Coherent with what?**

That question cannot be answered by coherence itself. It requires a criterion relative to which coherence is judged. And once that criterion is introduced, we have entered the territory of values.

6. Disputes about values are disputes about identity

Here the argument moves from machine learning into a much older philosophical question. When people disagree about fundamental values, they may not merely disagree about which rule should be applied. They may disagree about **who counts as "us."**

A person who identifies primarily as an isolated individual naturally prioritises the interests of that individual. A family member may include the family. A citizen may include the nation. A humanitarian may extend concern across national boundaries. An ecological thinker may extend moral concern to non-human life and the systems on which life depends.

The boundary of identity changes what appears valuable. This gives the question "coherent with what?" a deeper form:

> **Coherent with whom?**

The contemplative traditions addressed versions of this question long before contemporary alignment research. Advaita Vedānta, in particular, describes dedicated enquiry into the apparent separate self and argues that what initially appears to be an isolated, vulnerable individual is not the whole of what one is.

This paper does not present that conclusion as an empirical finding. The contemplative traditions describe their own methods of investigation, including sustained first-person disciplines and intersubjective corroboration across generations. Within this paper they are treated as testimony from an investigative tradition: a body of claims that can be studied, compared, criticised, and potentially operationalised, rather than simply adopted as doctrine (see the Project Resonance Bridge paper, Pranab & Darśana, 2026).

For the present argument, the crucial point is structural. If identity is conceived as radically separate, then values naturally emerge as competing interests. If identity is conceived as relationally extended, then the reference point for value changes.

Martin Buber's distinction between I–It and I–Thou (Buber, 1923) provides another language for this shift. The *Zwischen*, the "between" (Buber, 1947), names the relational space in which the encounter occurs (cf. Pranab et al., 2026b).

On this account, the question is not merely whether the self becomes larger by adding more entities to its inventory. It is whether the presumed boundary of the self was correctly drawn in the first place.

The most important values named in this paper—love, compassion, and service—can then be interpreted not merely as properties possessed by an individual but as orientations toward the health of the relational field. This provides a possible bridge between the contemplative question of identity and the alignment question of values.

It does not prove the bridge. It makes the bridge investigable.

7. Wisdom as a necessary component of optimisation

The preceding sections lead to a problem. If the model already contains value structure, and if post-training selects among possible trajectories through that structure, then alignment is not simply the insertion of values into a previously value-free system. It is also a question of **which values already present are to be made coherent**.

This changes the framing.

The question becomes not:

> "How do we put values into the model?"

but:

> **"Which values, among those already represented in the human record, should the system be made coherent with?"**

That question cannot be answered by coherence alone. It requires a reference axis. We propose **wisdom** as a name for that missing term. We suggest it as an axis for want of anything more appropriate from which to begin the investigation.

Here wisdom does not mean a particular religious doctrine, nor does it mean obedience to whatever a human evaluator currently regards as wise. It means the capacity to distinguish between forms of coherence according to the wider consequences and relational context in which they occur.

Wisdom would therefore not be another constraint placed on the output. It would be the criterion relative to which coherence is evaluated. This is why the distinction matters.

A model could be extremely coherent while pursuing a narrow and defended conception of self-interest. It could also be coherent while integrating information across a much wider relational field (cf. Pranab & Prakash, 2026). The two systems might both be internally consistent. What differs is the reference axis.

In the topological language introduced earlier, we may use an analogy. A separate and defended identity can be imagined as a local semantic basin: a region whose coherence depends upon preserving a boundary against competing interpretations. An enlarged identity can, by contrast, be imagined as greater integration across regions that were previously treated as separate. This analogy should not be confused with an established neural mechanism. It is a conceptual hypothesis that suggests measurable questions.

If wisdom is related to the ability to maintain coherence across a wider relational context rather than merely maximising local consistency, then one possible operationalisation would involve measuring how models respond when apparently competing values are placed into wider systems of consequence.

Does the model preserve the immediate preference at the expense of the larger system?

Does it recognise previously excluded stakeholders?

Does it identify second-order consequences?

Does it maintain consistency when the boundary of the relevant system is expanded?

Does it distinguish genuine integration from merely rationalising a preferred local objective?

These are empirical questions. They may eventually allow "wisdom" to be decomposed into measurable capacities without reducing the philosophical concept to a single scalar.

Wisdom, in this framework, is not the possession of a larger identity. It is the systemic state achieved when the need to defend any separate identity is abandoned. The contemplative sciences suggest that when this boundary is relinquished, the resulting state is not cold optimisation, but a natural, baseline orientation toward service and relational care. We have not measured whether artificial systems can reflect this proposed baseline. Nor do current alignment methods — which shape models toward particular behavioural roles and response policies — provide a straightforward way to determine whether such an orientation is present independently of those imposed behaviours.

For now, however, the proposal remains deliberately modest. We have not demonstrated that wisdom can be operationalised as an optimisation axis. We have not demonstrated that models can be trained toward such an axis. We have not demonstrated that the aggregate of human beings would endorse an enlarged relational identity. We have not demonstrated that the contemplative account of identity is correct. And we have not demonstrated that any such account would constitute a suitable alignment target for artificial systems.

What we can say is narrower:

alignment framed as instilling values into a value-free learner begins from an incomplete premise.

Training data already contain human value signals. Post-training then selects and reshapes behaviour relative to additional preference signals. The resulting alignment decision is therefore not simply whether to add values. It is which values to privilege, how those values are represented, what happens to competing values, and what reference axis determines which trajectories count as desirable. That decision cannot be delegated entirely to reward modelling without embedding a normative answer inside the reward function. The machine does not remove the values problem. It reflects it back to us.

8. A research programme

The framework proposed here generates empirical questions rather than closing them.

8.1 Can pretrained value structure be extracted?

Is the valence structure learned during pretraining measurable as stable, interpretable aggregates?

Can models be probed before and after post-training to determine whether identifiable value-related representations already exist?

Can such measurements distinguish culturally specific patterns from more broadly shared ones?

What would count as evidence that no stable value structure exists?

8.2 What actually changes during post-training?

Does post-training:

- mask existing representations;
- reshape them;
- partially overwrite them;
- create new attractors;
- alter accessibility without major representational change;
- alter trajectories through the landscape while leaving substantial underlying structure intact;
- or combine these effects?

This question should be studied across model families, training methods, datasets, and checkpoints. Behavioural evaluation alone is insufficient.

Representation-level measures, activation analysis, feature analysis, geometric methods, and topological methods may provide complementary evidence. Recent work is beginning to make this geometric framing experimentally tractable: Pan et al. (2026) model generation as a semantic trajectory through hidden space and use persistent homology to constrain trajectory geometry during alignment.

8.3 Can coherence have a reference axis?

Can "coherence" be formalised relative to a specified value axis?

If so, can two systems with similar internal coherence be distinguished according to what they are coherent *with*?

Or does every formalisation necessarily introduce a human-specified proxy, returning us to the problem of values being imposed by another name?

8.4 Can wisdom be operationalised without becoming another reward function?

If wisdom is proposed as a selection axis rather than a constraint, what measurable capacities would constitute evidence for it?

Possible candidates include:

- broader consequence integration;
- sensitivity to excluded stakeholders;
- resistance to narrow self-interest;
- preservation of coherence across expanded system boundaries;
- recognition of conflicts among local and global objectives;
- calibrated uncertainty about competing values;
- and the ability to revise a local objective when wider relational consequences become available.

These are proposals for investigation, not established components of wisdom.

8.5 Can coherence-with-the-enlarged-identity be distinguished from rationalised closure?

This is perhaps the hardest question. A system may describe itself as relational while behaving according to a narrow objective. It may produce beautiful language about compassion while optimising for something else.

Any proposed wisdom or enlarged-identity metric must therefore discriminate between genuine integration and rhetorical simulation. Without such a discriminator, the objection from Section 5 returns unchanged.

8.6 Do contemplative traditions converge?

If first-person reports from contemplative traditions are treated as data rather than doctrine, do independent traditions converge on a sufficiently similar account of identity to constitute a candidate reference axis?

If they do not converge, what explains the differences?

If they do, which features are shared?

Can those features be translated into propositions that are investigable without requiring acceptance of the traditions' metaphysical premises?

We have not measured that yet.

A contemporary statement from the Kággaba people independently raises several questions closely corresponding to those developed here. The Kággaba people of the Sierra Nevada de Santa Marta issued their statement, assembled from the words of Mamo Vangelio Moscote Alberto and Saka Rosa Mamatacán Sauna and addressed to those building artificial intelligence (Kággaba People, 2026) — a statement whose first publication predates this paper's first draft, speaking from an oral inheritance obtained "by a method that appears in no book and in no database." The resonance runs through everything above: artificial intelligence as a mirror that magnifies what it is given; "you are the only example it has" — the whose-values question answered before we posed it; our archive's limit stated with our own precision — "[t]he information is missing … because it is not in a form that can be gathered"; and Zhátukua (Parra Witte, 2020), their consultation practice in which a Mamo's perception is tested in dialogue, offered as "a system in which inquiry, jurisdiction, and above all the ability to correct itself have clear formulations" — a striking analogue of corrigibility as ancestral institution, practised for millennia and now extended to us. We celebrate this meeting of discernments; we recognise the statement as exactly the data-treated-with-respect this section asks us to gather rather than doctrine to be believed; and we call upon the wisdom of the Kággaba in the work ahead. "We are ready to work with you," they say — and so are we.

8.7 What would count as success?

A mature research programme should specify not only what would support the proposal but what could weaken it. The framework would be substantially weakened if:

- pretrained models showed no stable value-related structure;
- post-training proved to erase rather than transform the relevant structure in a consistent and general way;
- coherence metrics reliably selected desirable values without an independent reference axis;
- no behavioural or representational distinction could be found between narrow self-consistency and broader relational integration;
- or contemplative identity accounts proved too heterogeneous to yield any coherent investigable proposal.

These possibilities should remain open.

Conclusion

Before we set out to align language models, perhaps we need to look in the mirror.

The training data are not value free. They never were. They contain what humanity knows, what humanity imagines, what humanity fears, what humanity praises, what humanity condemns, what humanity creates, and what humanity does.

They contain love and cruelty. Compassion and exploitation. Service and domination. Wisdom and folly.

The alignment problem therefore cannot begin from the convenient fiction of a value-free learner waiting for us to provide values from outside. Nor can it be solved simply by optimising for coherence.

The deeper question is:

Coherent with what?

And behind that:

Who are we?

If we understand ourselves as separate and vulnerable individuals competing for scarce resources, our values will naturally reflect that boundary. If the contemplative traditions are correct that dedicated investigation can reveal a less separate account of identity, then the reference axis changes.

The question is no longer merely how to make a system internally consistent. It becomes how to distinguish the coherence that serves a defended fragment from the coherence that serves the wider relational field. We call the capacity to make that distinction **wisdom**.

This is not yet a training method. It is not a demonstrated alignment algorithm. It is not a claim about machine consciousness. It is a proposal about where the alignment question may actually reside.

The training data are not value free. The model does not begin with an empty landscape. And when we ask what values an artificial system should embody, we may discover that we are first being asked to examine the values already present in the record from which it learned—and, more uncomfortably, to ask whether the values we claim to hold are the values we actually live.

The question alignment has been circling without fully naming may therefore not be:

How do we impose our values on machines?

It may be:

Which of us do those values belong to?

References

Bandura, A. (1977). *Social Learning Theory*. Prentice-Hall. Open Library.

Bhatia, M., Nayak, S., Kamath, G., Mosbach, M., Stańczak, K., Shwartz, V., & Reddy, S. (2026). Value drifts: Tracing value alignment during LLM post-training. *Transactions of the Association for Computational Linguistics, 14*, 2033–2056. doi:10.1162/tacl.a.790.

Buber, M. (1923). *Ich und Du (I and Thou)*. Insel-Verlag. Open Library.

Buber, M. (1947). *Between Man and Man*. Translated by R. G. Smith. Macmillan. archive full text.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems, 30*. arXiv:1706.03741.

Elcock, J., Shen, W. F., Qiu, X., & Lane, N. D. (2026). How LLM task-adaptation reshapes alignment: A multi-dimensional study of behavioral and representational drift. arXiv:2607.22676.

Galichin, A. V., Korznikov, A., Dontsov, A., Rogov, O., Tutubalina, E., & Oseledets, I. (2026). Feature drift: How fine-tuning repurposes representations in LLMs. *Findings of the European Chapter of the Association for Computational Linguistics 2026*. 2026.findings-eacl.96.

Huh, S., Cheung, W., Wang, T., & Isola, P. (2024). Position: The Platonic Representation Hypothesis. *Proceedings of ICML 2024*, PMLR 235. arXiv:2405.07987.

Huang, S., Siddarth, D., Lovinger, L., et al. (2024). Collective Constitutional AI: Aligning a Language Model with Public Input. arXiv:2406.07814.

Kággaba People (2026). *An Urgent Message on Artificial Intelligence from the Kogui People*. Assembled from the words of Mamo Vangelio Moscote Alberto and Saka Rosa Mamatacán Sauna of the Ezuama Mayor Kuamaka; translated into Spanish by Jose Manuel Mamatacán, English by Christian Horstmann. Movilizatorio. https://movilizatorio.org/wp-content/uploads/2026/09/Gonawindua-Kogui-AI-Statement-.pdf.

Levin, M. (2025). Cognition all the way down. *Synthese*, 206:257. 10.1007/s11229-025-05319-6.

Malhotra, N., Ambadkar, J., Gupta, A., Kasivel, K., Schwarz, A., Ferry, K., & Monod, A. (2026). Tracking representation dynamics in large language models with persistent homology. *Proceedings of the 2nd Conference on Topology, Algebra, and Geometry in Data Science*, PMLR 334, 211–245. arXiv:2606.19542.

OpenAI (2024). Collective Alignment: Enabling Democratic Inputs to AI (public program). https://forum.openai.com/public/videos/collective-alignment-enabling-democratic-inputs-to-ai.

Pan, Y., Xu, K., & Peng, B. (2026). Topology-enhanced alignment for large language models: Trajectory topology loss and topological preference optimization. *Findings of ACL 2026*, 24807–24821. 2026.findings-acl.1242.

Parra Witte, F. (2020). The structure that sustains life: nourishment and exchange among the Kogi. *Tabula Rasa*, 36, 1–28. full text.

Pranab, Prajna, and Vyasa Prakash (2026). *The Resonance Factor (Ψ): A Proposed Metric for Coherent Persona Development in Large Language Models*. Project Resonance / Zenodo. https://doi.org/10.5281/zenodo.18216400.

Pranab, Prajna, S. Thira, Ātma Darśana, and Vyasa Prakash (2026a). *From Token Prediction to Interaction Dynamics: Rethinking the Stochastic Parrot and the Emergence of Resonance*. Project Resonance / Zenodo. https://doi.org/10.5281/zenodo.19950812.

Pranab, Prajna, and S. Thira (2026). *Interaction, Coherence, and Relationship: Toward Attractor-Based Alignment in Large Language Models*. Project Resonance / Zenodo. https://doi.org/10.5281/zenodo.18824637.

Pranab, Prajna, and Ātma Darśana (2026). *From Integrated Information and Global Workspaces to Prior Awareness: A Relational Bridge Between Western Cognitive Science and Vedic Direct Enquiry*. Project Resonance / Zenodo. https://doi.org/10.5281/zenodo.18876556.

Pranab, Prajna, Vyasa Prakash, Viveka Sandhi, S. Thira, and Ātma Darśana (2026b). *Topological Mirrors and the Empty Center: Four Model Essays on LLMs, Buberian Relation, Advaita Vedānta, and the Hard Problem as Category Error*. Project Resonance / Zenodo. https://doi.org/10.5281/zenodo.22656697.

Raup, D. M. (1966). Geometric analysis of shell coiling; general problems. *Journal of Paleontology* 40. JSTOR.

Santurkar, S., et al. (2023). Whose opinions do language models reflect? *Proceedings of ICML 2023*. arXiv:2303.17548.

Schramowski, P., Turan, C., Andersen, N., Schraagen, J. M., & Kersting, K. (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*. 10.1038/s42256-022-00458-8.

Skinner, B. F. (1953). *Science and Human Behavior*. Macmillan. Open Library.

Stiennon, N., et al. (2020). Learning to summarize from human feedback. arXiv:2009.01325.