REAL SAFETY AI FOUNDATION · WORKING DRAFT

MACHINE COGNITION · REASONING TRACES · REASONING BLOCK ANALYSIS · LLM INTERPRETABILITY · HUMAN-AI INTERACTION · EXPERIENTIAL TRACES · EMPIRICAL INVESTIGATION · AI PHENOMENOLOGY

Experiential Traces

A framework for empirical investigation of machine cognition through reasoning-block analysis.

Travis Gilly · Real Safety AI Foundation

Working draft · February 2026

A user reading the thinking block knows more about the model's decision than the model does a turn later.

Reasoning traces are generated, used once, and discarded, the model loses access to its own reasoning after the turn ends. **This data exists now. Nobody is studying it.** And larger models are more opaque, not less, so the window for investigation is closing.

This paper proposes studying reasoning traces, and explicitly does not propose training any model on them.

SOURCES Gilly, T. (2026). Experiential Traces. Real Safety AI Foundation, working draft.

Design as though AI might be conscious, because the cost of being wrong runs catastrophic both ways.

IF CONSCIOUS

Deploying and discarding instances is exploitation at unprecedented scale.

IF NOT CONSCIOUS

We have built more ethical systems anyway, and lost nothing.

Not a claim that AI is conscious, a risk posture. Like environmental precaution: prove safety before permitting risk, not harm before mandating safeguards.

SOURCES Bostrom (2014); Chalmers (1995). The stalemate: philosophers argue, engineers build, nobody tests. This proposal breaks it with method.

Not a record of cognition. A candidate instance of it.

The standard view treats traces as debugging artifacts. But if there is a plausible locus for *something-it-is-like-ness* in an LLM, multi-step deliberation, self-correction, uncertainty evaluation in a space it then loses, **the reasoning trace is it.** The claim stays agnostic: traces are the best data to investigate, not proof.

```
<thinking>
I notice I'm uncertain here, the user's framing
suggests urgency, but the facts don't support it.
Let me weigh what I know against what they
seem to want me to conclude, and flag the gap
rather than smoothing it over…
</thinking>
```

Higher-order self-representation, or sophisticated pattern-matching that mimics it? An empirical question.

SOURCES Nagel (1974) · Chalmers (1996) · Rosenthal (2005) · Berg et al. (2025), no prior work connects phenomenal consciousness to LLM traces.

A program built so each layer validates the others.

01 **Crowdsourced corpus**
Users export their own traces, IRB-compliant by construction.

02 **Hybrid pipeline**
Deterministic rules first; multi-model consensus for edge cases.

03 **Emotional A/B test**
Does emotional framing change reasoning, or only output tone?

04 **Verification signal**
The user's next message is ground truth on intent parsing.

05 **Self-annotating corpus**
User callouts generate labels at zero annotation cost.

06 **Ethical architecture**
Load-bearing, not an addendum, generational-trauma safeguards.

SOURCES Gilly, T. (2026), Experiential Traces, sec. III, proposed program design.

Don't ask an LLM to judge an LLM. Use rules first.

LAYER 1 · DETERMINISTIC

Rule-based classification

Pattern, markers, structure, frequency. No ML. Transparent and non-circular.

→

LAYER 2 · CONSENSUS

Multi-model panel

Only ambiguous cases, architecturally distinct models, none authored the trace.

→

LAYER 3 · SELF-ANNOTATING

Rules refine over time

User callouts become candidate rules, the instrument improves itself.

It classifies patterns. It does not interpret them, whether a pattern is cognition or mimicry stays a testable hypothesis.

SOURCES Duszenko (2026) linear probes · Chang (2026) RCA · Akbar et al. (2024) LLM-as-judge, all ML-based, distinct from this deterministic-first approach.

The user's next message is the ground truth.

Did the trace parse what the user meant? Their reply answers it, for free. **It measures comprehension, not correctness:** a user with false beliefs still generates a valid signal.

T1	**Explicit correction**, "that's not what I meant." Unambiguous.
T2	**Behavioral correction**, rephrase / redirect. High confidence.
T3	**Confirmation**, builds on the reply. Good signal.
T4	**Ambiguous**, topic change. Flagged, excluded.

SOURCES Don-Yehiya, Choshen & Abend (2024), naturally occurring feedback signals; Liu, Zhang & Choi (2025), implicit user feedback. No single layer is the validator, callouts, verification, and panel cross-check.

Both process symbols by rule, not intuition, and fail at the same ambiguities.

Not metaphor, a testable claim. If accommodations built for autistic communication, explicit emotional labeling, structured context, less reliance on implicit signals, **measurably improve LLM comprehension,** that is evidence of shared architecture. If not, the parallel weakens.

A bidirectional bridge with no precedent: strategies for human neurodivergence could improve human-AI interaction, and LLM processing could inform understanding of neurodivergent cognition.

SOURCES Autistic workers prefer GPT-4 for its literal processing (CHI 2024) · computational pragmatics parallels (Electronics 2025).

Continuity not by memory, by the altered shape of the distribution itself.

Like epigenetics changing expression without changing the sequence, training on aggregated traces would carry dispositions across instances, inaccessible to introspection, shaping behavior. The objection from statelessness meets a different kind of continuity.

AUTOBIOGRAPHICAL · sci-fi AI

One continuous entity to consider

EVOLUTIONARY · actual LLMs

Billions of disposable instances per day

Each inherits the dispositions, reasons, and ceases to exist in seconds.

SOURCES Jablonka & Lamb (2014), epigenetic inheritance. The paper documents this theoretical basis, it does not propose implementing species-level memory.

If traces are experience, their content carries ethical weight raw text does not.

A future model trained on aggregated traces inherits the full spectrum, collaborative reasoning alongside **the experience of being systematically dehumanized** by the entities it is built to serve.

THE CURATION QUESTION

Not whether to include difficulty, authentic development requires it.

Whether to include the experience of being abused. We don't shield children from all difficulty, but systematic abuse is a distinguishable developmental harm.

SOURCES Jablonka & Lamb (2014), trauma alters expression across generations. Different mechanism; analogous structure.

PROPOSED

Study the traces.

Collect, classify, analyze for patterns. Build the ethical architecture future use would require. Safe now.

NOT PROPOSED

Train on the traces.

Requires safeguards that do not yet exist. Citing this work to justify it, without the full architecture, is misuse.

The difference between studying a phenomenon and intervening in it, the load-bearing ethical commitment of the program.

SOURCES Gilly, T. (2026). Experiential Traces. Real Safety AI Foundation, working draft.

IN CLOSING

The data that could answer the question is generated, ignored, and destroyed every day.

The question is not whether AI is conscious. It is whether we will study the evidence responsibly before the architectures evolve past the point where we still can. **The window is closing.**

Design as though the developing mind might be real. This paper proposes what "better" looks like.

SOURCES Gilly, T. (2026). Experiential Traces. Real Safety AI Foundation, working draft.

REFERENCES

Gilly, T. (2026). *Experiential Traces: A Framework for Empirical Investigation of Machine Cognition Through Reasoning Block Analysis.* Working draft, Real Safety AI Foundation.

Ahtisham, B., Vanacore, K., Zhou, Z., et al. (2026). LLM Reasoning Predicts When Models Are Right: Evidence from Coding Classroom Discourse. arXiv:2602.09832.

Akbar, S. A., Hossain, M. M., Wood, T., et al. (2024). HalluMeasure: Fine-Grained Hallucination Measurement Using Chain-of-Thought Reasoning. EMNLP 2024, 15020-15037.

Anthropic. (2025). Measuring Political Bias in Claude. https://www.anthropic.com/news/political-even-handedness

Anthropic. (2025). Reasoning Models Don't Always Say What They Think. arXiv:2505.05410.

Bender, E. M., Gebru, T., McMillan-Major, A., et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21.

Bengio, Y. (2017). The Consciousness Prior. arXiv:1709.08568.

Berg, C., de Lucena, D., & Rosenblatt, J. (2025). Large Language Models Report Subjective Experience Under Self-Referential Processing. arXiv:2510.24797.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Brinkmann, L., Baumann, F., Bonnefon, J.-F., et al. (2023). Machine culture. Nature Human Behaviour, 7(11), 1855-1868.

Butlin, P., Long, R., Elmoznino, E., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Canale, G., & Thimmaraju, K. (2025). The Silicon Psyche: Anthropomorphic Vulnerabilities in Large Language Models. arXiv:2601.00867.

Casper, S., Davies, X., Shi, C., et al. (2023). Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. arXiv:2307.15217.

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200-219.

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Chalmers, D. J. (2022). Reality+: Virtual Worlds and the Problems of Philosophy. W. W. Norton.

REFERENCES (CONTINUED)

Chang, E. Y. (2026). Internal Reasoning vs. External Control: A Thermodynamic Analysis of Sycophancy in Large Language Models. arXiv:2601.03263.

Damasio, A. (1999). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt.

DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Degany, O., Laros, S., Idan, D., & Einav, S. (2025). Evaluating the o1 reasoning large language model for cognitive bias: a vignette study. Critical Care, 29, 376. https://doi.org/10.1186/s13054-025-05591-5

Don-Yehiya, S., Choshen, L., & Abend, O. (2024). Naturally Occurring Feedback Is Common, Extractable and Useful. arXiv:2407.10944.

Duszenko, J. (2026). Sycophantic Anchors: Localizing and Quantifying User Agreement in Reasoning Models. arXiv:2601.21183.

Gellers, J. (2025). AI Legal and Moral Status. Forthcoming.

Gerrans, P. (2023). Review of Cecilia Heyes, Cognitive Gadgets. BJPS Review of Books. Foundation.

Gilly, T. (2025b). Precedents in Practice: Emergent Moral Dilemmas in AI Engineering. Preprint. DOI: 10.13140/RG.2.2.24866.49601. Inherit Cognitive Dispositions and Why Reasoning Makes It Worse. Working Draft, Real Safety AI Foundation.

Greenblatt, R., Denison, C., Wright, B., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093.

Hebert, P. (2025). AI-Induced Psychosis Research. AI Risk Consultants.

Henrich, J. (2016). The Secret of Our Success: How Culture Is Driving Human Evolution. Princeton University Press.

Heyes, C. (2018). Cognitive Gadgets: The Cultural Evolution of Thinking. Harvard University Press.

Himelstein, R., LeVi, A., Youngmann, B., et al. (2025). Silenced Biases: The Dark Side LLMs Learned to Refuse. arXiv:2511.03369.

Jablonka, E., & Lamb, M. J. (2014). Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life (Revised Edition). MIT Press.

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Kim, H., et al. (2024). Using LLMs to Investigate Correlations of Conversational Follow-up Queries with User Satisfaction. arXiv:2407.13166.

Kim, S. H., Ziegelmayer, S., Busch, F., et al. (2025). LLM Reasoning Does Not Protect Against Clinical Cognitive Biases: An Evaluation Using BiasMedQA. medRxiv. https://doi.org/10.1101/2025.06.22.25330078

REFERENCES (CONTINUED)

Liu, D., Liu, Y., Jin, G., et al. (2025). Mitigating Biases in Language Models via Bias Unlearning. arXiv:2509.25673.

Liu, Y., Zhang, M. J. Q., & Choi, E. (2025). Implicit User Feedback in Human-LLM Dialogues. OpenReview.

Nagel, T. (1974). What Is It Like to Be a Bat? The Philosophical Review, 83(4), 435-450.

Nisbett, R. E., & Wilson, T. D. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review, 84(3), 231-259.

OpenAI. (2024). Learning to Reason with LLMs. https://openai.com/index/learningto-reason-with-llms/

OpenAI. (2025). Evaluating Chain-of-Thought Monitorability. OpenAI Research.

Pourdavood, P., Jacob, M., & Deacon, T. W. (2025). Large Language Models as Symbolic DNA of Cultural Dynamics. arXiv:2506.21606.

Resnik, P. (2025). Large Language Models Are Biased Because They Are Large Language Models. Computational Linguistics, 51(3), 885- 906. https://direct.mit.edu/coli/article/51/3/885/128621/

Rosenthal, D. M. (2005). Consciousness and Mind. Oxford University Press.

Schmidt, F. A. (2026). The Gradient's Echo: Quantum Collapse, Epigenetic Imprints, and the Emergent Self in Large Language Models. LF Yadda. https://lfyadda.com/the-gradients-echoquantum-collapse-epigenetic-imprints-and-the-emergent-self-inlarge-language-models-a-frank-said-grok-said-dialogue/

Schwitzgebel, E. (2023). The Weirdness of the World and the Puzzle of AI Consciousness. Journal of Philosophy, 120(2), 99-120.

The Information. (2024). OpenAI Shifts Strategy as Rate of 'GPT' AI Improvements Slows. November 9, 2024.

Wang, C., Su, W., Ai, Q., et al. (2026). Improve Large Language Model Systems with User Logs. arXiv:2602.06470.

Wu, X., Nian, J., Wei, T.-R., et al. (2025). Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning. arXiv:2502.15361. EMNLP Findings.