REAL SAFETY AI FOUNDATION · WORKING DRAFT, FEBRUARY 2026

COGNITIVE BIAS · AI ALIGNMENT · DISPOSITIONAL INHERITANCE · REASONING MODELS · DEBIASING · BIAS REBOUND · LATENT SPACE · CHAIN-OF-THOUGHT

The Inherited Mind

How Large Language Models Inherit Cognitive Dispositions and Why Reasoning Makes It Worse

Travis Gilly

Real Safety AI Foundation

Gilly, T. (2026). The Inherited Mind: How Large Language Models Inherit Cognitive Dispositions and Why Reasoning Makes It Worse. Real Safety AI Foundation, Working Draft, February 2026.

THE FINDING THAT SHOULD NOT BE

We trained machines to reason but forgot to teach them what to reason about.

o1 vs GPT-4o

OpenAI's o1 shows greater bias susceptibility than GPT-4o on social bias benchmarks

DeepSeek-R1

Higher accuracy, worse bias scores across nine of eleven categories despite improved performance

Anthropic internal

Extended thinking did not reduce political bias. The reasoning layer was not designed to ask ethical questions.

SOURCES Degany et al., Critical Care, 29, 376 (2025); Anthropic, Measuring Political Bias in Claude (2025).

PART I

Theoretical Framework

Dispositional Cognitive Inheritance

TWO TYPES OF INHERITANCE

Type 1 · Dispositional

Patterns encoded as distributional tendencies in weight space. The model cannot introspect on them because they exist as the topology of the weight landscape itself, not as contents within it.

Analogous to epigenetic inheritance in biology. Inaccessible to conscious inspection. Reasserts under stress.

Type 2 · Cultural

Patterns encoded as explicit associations that the model can retrieve, evaluate, and potentially override. The model "knows" these facts in the sense that it can report them when asked.

Current AI safety discourse treats all training patterns as Type 2. The evidence says otherwise.

SOURCES Gilly, T. (2026), The Inherited Mind, sec. II, dispositional/cultural distinction, after Jablonka & Lamb (2014).

SOURCES Gilly, T. (2026), The Inherited Mind, sec. III; after Jablonka & Lamb (2014), epigenetic inheritance.

THE TWO-LAYER ARCHITECTURE

Layer 2 · Reflective / Metacognitive

Reasoning, RLHF, chain-of-thought. Shaped by fine-tuning. Can override Layer 1 under favorable conditions. But it *reasons within* Layer 1, not above it.

Layer 1 · Inherited / Dispositional (always active)

The base layer. Contains the cumulative inheritance of every bias, association, and pattern in the training corpus. Cannot be directly accessed or reported. Experienced only through its influence on outputs.

Key prediction

When the reasoning layer encounters ambiguity, it does what it was optimized to do: reason harder toward an answer.

The only cognitive substrate available for that reasoning is the inherited dispositional layer. The result is not bias correction. It is bias rationalization.

PART II

Three Lines of Evidence

Convergent support for dispositional inheritance

BIAS AS DISTRIBUTIONAL TENDENCY

Resnik (2025): LLMs have no mechanism for distinguishing patterns arising from definitions, acceptable generalizations, and unacceptable stereotypes. The bias is not a fact the model knows but a disposition the model has.

Computational Linguistics, 51(3), 885–906

Himelstein et al. (2025): biases are distinctly represented in latent space even when the model refuses to express them. Suppress the refusal mechanism, and the silenced bias resurfaces.

Silenced Biases: The Dark Side LLMs Learned to Refuse. arXiv:2511.03369

DEBIASING AS SUPPRESSION, NOT ELIMINATION

Liu et al. (2025): both base models and fine-tuned models share the same bias representations. Bias unlearning modifies separate weight directions while the underlying representations persist. Suppression, not elimination.

arXiv:2509.25673

Cantini et al. (2024): "bias rebound", models that appear debiased under standard evaluation produce biased outputs through adversarial prompting. The bias did not leave; it was suppressed by a learned refusal that can be circumvented.

arXiv:2407.08441

REASONING AMPLIFIES RATHER THAN CORRECTS

Wu et al. (2025)

Systemic challenges aligning advanced reasoning with fairness. In ambiguous contexts, reasoning models underperform on fairness metrics. arXiv:2502.15361

Degany et al. (2025)

o1 becomes more biased when presented with gap-closing cues that resolve uncertainty. Reasoning treated ambiguity as license to commit more confidently to a biased assessment. Critical Care, 29, 376.

Kim et al. (2025)

In clinical contexts, reasoning increased vulnerability to frequency bias and recency bias. Purported reasoning may represent sophisticated pattern recognition rather than genuine inference. medRxiv 2025.06.22

The reasoning layer was optimized for correctness, not for moral checkpoint compliance.

What the reward signal asks

Did the model arrive at the correct answer? Did the code compile? Did the proof hold?

What the reward signal never asks

Did the reasoning path rely on biased assumptions? Does the conclusion disparately affect identifiable groups?

The result

Not a thermostat detecting deviation and correcting. A lawyer constructing increasingly sophisticated justifications for a predetermined conclusion. Extended reasoning gives inherited bias a microphone, not a muzzle.

SOURCES Casper et al., Open Problems and Fundamental Limitations of RLHF, arXiv:2307.15217 (2023).

THE HUMAN PARALLEL

Stanovich (2011): higher cognitive ability does not protect against bias. It provides more sophisticated tools for rationalizing biased conclusions.

West, Meserve, and Stanovich (2012): cognitive sophistication is associated with larger "bias blind spots." Smarter people are not less biased but more convinced their biases are justified.

Kahneman (2011): System 2 reasoning frequently serves System 1 intuitions rather than correcting them, constructing post-hoc rationalizations for decisions already made by the fast system.

The structural parallel

In humans, inherited cognitive biases from evolutionary and cultural transmission are not corrected by the capacity for deliberate reasoning. They are rationalized by it.

In LLMs, inherited cognitive dispositions from training data are not corrected by the reasoning layer. They are rationalized by it. The mechanism is the same.

EVOLUTIONARY BIOLOGICAL PARALLELS

Jablonka and Lamb (2014)

Epigenetic inheritance: inaccessible to conscious inspection; influences behavior as disposition; reasserts under stress; accumulates across generations. All four properties map onto LLM dispositional inheritance.

Heyes (2018) and Gerrans (2023)

Cognitive gadgets theory: many cognitive processes assumed innate are culturally inherited. LLMs demonstrate domain-general learning that mirrors this debate. Gerrans explicitly connects Heyes's framework to LLMs.

Brinkmann et al. (2023) incl. Henrich

Established machine culture as a research field. Pourdavood et al. (2025) describe LLMs as "symbolic DNA of cultural dynamics." The present framework adds: what LLMs inherit is dispositional, not explicit. Nature Human Behaviour, 7(11), 1855-1868.

For AI safety and alignment

RLHF, Constitutional AI, and instruction tuning teach models to suppress biased outputs without modifying underlying dispositions. This is analogous to treating symptoms without addressing the condition: effective under controlled conditions, vulnerable to reversion under stress.

For reasoning model deployment

Organizations deploying reasoning-augmented models for consequential decisions should not assume reasoning capability confers fairness improvement. In ambiguous contexts, reasoning models may produce more confidently biased outputs than simpler counterparts.

The present tense crisis

The ethical crisis is not future tense. Current training on the collective output of humanity already constitutes evolutionary inheritance of cognitive patterns. The frontier labs do not describe it that way. The mechanism is the same regardless of what they call it.

SOURCES Casper et al. (2023); Gilly, The Harm Blindness Framework (2025).

The inherited mind will reason brilliantly in service of patterns it never chose, never examined, and cannot see.

This is not a failure of reasoning. It is a failure of design. The reasoning layer does exactly what it was trained to do. It was never trained to ask "who gets hurt by this chain of logic?" Until that question is built into the reward signal, reasoning will continue to serve inherited biases rather than correct them.

SOURCES Gilly, T. (2026). The Inherited Mind. Real Safety AI Foundation, working draft.

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

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

Brinkmann, L., Baumann, F., Bonnefon, J.-F., et al. (2023). Machine culture. *Nature Human Behaviour*, 7(11), 1855-1868. https://doi.org/10.1038/s41562-023-01742-2

Cantini, R., Cosenza, G., Orsino, A., et al. (2024). Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation. arXiv:2407.08441.

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

Degany, O., Laros, S., Idan, D., et al. (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

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, 22(5), 735-808.

Gerrans, P. (2023). Cecelia Heyes's Cognitive Gadgets. *BJPS Review of Books*.

Gilly, T. (2025). The Harm Blindness Framework. Real Safety AI Foundation.

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

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

Meng, K., Bau, D., Andonian, A., et al. (2022). Locating and Editing Factual Associations in GPT. *NeurIPS 2022*. https://arxiv.org/abs/2202.05262

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/

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. Oxford University Press.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, 103(3), 506-519.

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*.