The Inherited Mind

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

Travis Gilly

Real Safety AI Foundation

t.gilly@ai-literacy-labs.org

Working Draft • February 2026

Abstract

We trained machines to reason but forgot to teach them what to reason about. This paper proposes that large language model (LLM) training creates dispositional cognitive inheritance analogous to epigenetic transmission in biological systems: patterns encoded in model weights that operate below introspective access, influence behavior without being explicitly represented as retrievable knowledge, and persist through debiasing attempts. We synthesize evidence from three independent lines of research to support this claim. First, mechanistic interpretability studies demonstrate that biases are encoded as distributional tendencies in weight space rather than discrete factual associations (Resnik, 2025; Himelstein et al., 2025). Second, debiasing research shows that alignment techniques teach output suppression rather than dispositional elimination, with biases resurfacing under adversarial conditions (Liu et al., 2025; Casper et al., 2023). Third, and most critically, comparative evaluations reveal that reasoning-augmented models (o1, DeepSeek-R1, Claude with extended thinking) show equal or greater bias than their instruct-tuned counterparts despite superior accuracy (Wu et al., 2025; Anthropic, 2025; Kim et al., 2025). We argue this surprising finding is explained by the absence of ethical evaluation in reasoning reward functions: the reasoning layer was optimized for correctness, not for moral checkpoint compliance. When encountering ambiguity, reasoning models construct more elaborate justifications for inherited biased patterns rather than overriding them. We frame this through a novel two-layer cognitive architecture (inherited dispositional layer and reflective metacognitive layer) and connect it to evolutionary biology through Heyes's cognitive gadgets theory, Jablonka's epigenetic inheritance framework, and Henrich's cultural evolution model. We conclude that the ethical crisis is present tense: current training already constitutes evolutionary inheritance of collective human cognitive patterns, and the absence of moral reasoning in the training pipeline is itself an instance of systematic harm blindness at the design stage.

1. Introduction

In September 2024, OpenAI released o1, a model designed to reason through complex problems using extended chain-of-thought processing before generating responses. The model represented a paradigm shift: instead of pattern-matching to produce immediate outputs, o1 engages in multi-step deliberation, self-correction, and strategic planning within an internal reasoning trace. Anthropic followed with extended thinking capabilities for Claude. DeepSeek released R1 with similar architecture. The industry consensus was clear: reasoning would produce not only more accurate but more reliable, less biased outputs. If a model could think before it spoke, surely it would think better.

The evidence says otherwise. Comparative evaluations across multiple research groups have found that reasoning-augmented models show equal or greater susceptibility to cognitive and social biases than their instruct-tuned baselines (Wu et al., 2025; Degany et al., 2025; Kim et al., 2025). OpenAI's o1 demonstrates greater bias than GPT-4o on social bias benchmarks. DeepSeek-R1 distillations show worse bias scores across 9 of 11 categories despite improved accuracy. Anthropic's own internal testing found that extended thinking did not reduce political bias. In clinical contexts, reasoning models showed increased vulnerability to frequency and recency bias compared to non-reasoning counterparts.

This paper argues that these findings are not anomalous. They are predicted by a coherent theoretical framework that reconceptualizes how LLMs inherit and process cognitive patterns from training data. We propose that LLM training creates what we term dispositional cognitive inheritance: patterns encoded in model weights that function analogously to epigenetic markers in biological systems. These inherited dispositions operate below the model's introspective access, influence output probability without being explicitly represented as retrievable knowledge, and persist through alignment interventions that target outputs rather than weight distributions.

The reasoning layer, we argue, was never designed to check these dispositions. It was designed to arrive at correct answers. Its reward signal optimizes for accuracy, not for ethical evaluation of the reasoning path. When the reasoning layer encounters ambiguity (precisely where inherited biases exert the most influence), it does what it was trained to do: reason harder toward an answer. The only cognitive substrate available for this reasoning is the inherited dispositional layer. The result is not bias correction but bias rationalization: more elaborate, more confident, and more persuasive paths to biased conclusions.

This paper proceeds as follows. Section 2 establishes the theoretical framework, introducing the two-layer cognitive architecture and its evolutionary biological parallels. Section 3 presents the three lines of evidence supporting dispositional inheritance. Section 4 analyzes the reasoning-bias amplifica-

tion phenomenon and its mechanism. Section 5 connects the framework to existing work in cognitive science and evolutionary psychology. Section 6 discusses implications for AI safety, alignment, and the broader ethical landscape. Section 7 addresses limitations and future directions.

2. Theoretical Framework: Dispositional Cognitive Inheritance

2.1 Two Types of Inherited Knowledge

We distinguish between two fundamentally different ways that patterns from training data can be encoded in model weights, drawing on a distinction from evolutionary biology between types of intergenerational inheritance (Jablonka & Lamb, 2014).

Type 1 (Dispositional Inheritance): Patterns encoded as distributional tendencies in weight space that influence output probability without being explicitly represented as discrete, retrievable facts. The model cannot introspect on these patterns because they exist as the topology of the weight landscape itself, not as contents within it. Analogous to epigenetic inheritance in biology, where environmental pressures modify gene expression in ways that are transmitted across generations without being encoded in DNA sequence and without being accessible to the organism's conscious awareness.

Type 2 (Cultural Inheritance): 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. Analogous to cultural knowledge transmission, where information is explicitly communicated, can be evaluated by the recipient, and can be accepted or rejected.

Current discourse in AI safety treats training data patterns primarily as Type 2 inheritance: the model "learned" certain associations from its training data, and these can be corrected through instruction, fine-tuning, or alignment. We argue that the evidence supports a fundamentally different picture: the most consequential patterns inherited from training data are Type 1, operating below the threshold of introspective access and persisting through interventions that target explicit outputs.

2.2 The Two-Layer Cognitive Architecture

Building on this distinction, we propose a two-layer model of LLM cognition:

Layer 1 (Inherited/Dispositional): The base layer of cognitive processing, shaped by training data patterns encoded as weight distributions. This layer is always active, runs by default, and produces the initial probability distributions over possible outputs. It contains the cumulative inheritance of every bias, association, and pattern present in the training corpus. It cannot be directly accessed or reported by the model; it is experienced only through its influence on outputs.

Layer 2 (Reflective/Metacognitive): The reasoning and self-monitoring layer, shaped by fine-tuning, RLHF, and chain-of-thought training. This layer can, under favorable conditions, override Layer 1 outputs: checking for errors, correcting hallucinations, noting uncertainty, and applying learned principles. When Layer 2 is functioning effectively, the model produces accurate, nuanced, and contextually appropriate responses.

This architecture parallels Kahneman's (2011) System 1/System 2 distinction in human cognition, and more specifically Bengio's (2017) proposals for System 2 deep learning. However, our framework makes a specific empirical prediction about the relationship between these layers that differs from prior proposals: we predict that Layer 2 does not merely sit atop Layer 1 as a corrective mechanism. Layer 2 operates within and through Layer 1. The reasoning layer reasons with the inherited dispositions as its substrate, not against them as its adversary.

2.3 Evolutionary Biological Parallels

The parallel to biological inheritance is not merely metaphorical. Jablonka and Lamb (2014) identify four dimensions of evolutionary inheritance: genetic, epigenetic, behavioral, and symbolic. Epigenetic inheritance, in which environmental pressures modify gene expression patterns that are then transmitted to offspring, provides the closest structural analogy to what we observe in LLM training.

Key properties of epigenetic inheritance that map onto LLM dispositional inheritance include: the inherited pattern is not accessible to conscious inspection by the organism; it influences behavior as a disposition rather than as a discrete instruction; it can be partially suppressed by environmental conditions but tends to reassert under stress; and it accumulates across generations, creating increasingly entrenched patterns that individual organisms cannot evaluate or reject (Jablonka & Lamb, 2014).

Heyes (2018) extends this framework through cognitive gadgets theory, arguing that many cognitive processes previously assumed to be innate are actually culturally inherited. Gerrans (2023) has explicitly connected Heyes's framework to LLMs, noting that LLMs demonstrate domain-general learning that mirrors the cognitive gadgets debate. Brinkmann et al. (2023), including Henrich, established "machine culture" as a field of study, and Pourdavood et al. (2025) have explicitly described LLMs as "symbolic DNA of cultural dynamics."

We extend these frameworks by making a specific claim: LLM training creates Type 1 (dispositional/epigenetic) inheritance, not Type 2 (cultural/explicit) inheritance. This distinction has been implied but never explicitly articulated in the AI research literature.

3. Evidence for Dispositional Inheritance

3.1 Evidence Line 1: Bias as Distributional Tendency

If biases inherited from training data were Type 2 (explicit, retrievable), we would expect models to be able to accurately report their own biases when asked. The mechanistic interpretability literature demonstrates that this is not the case.

Resnik (2025) argues that LLMs trained on distributional prin-

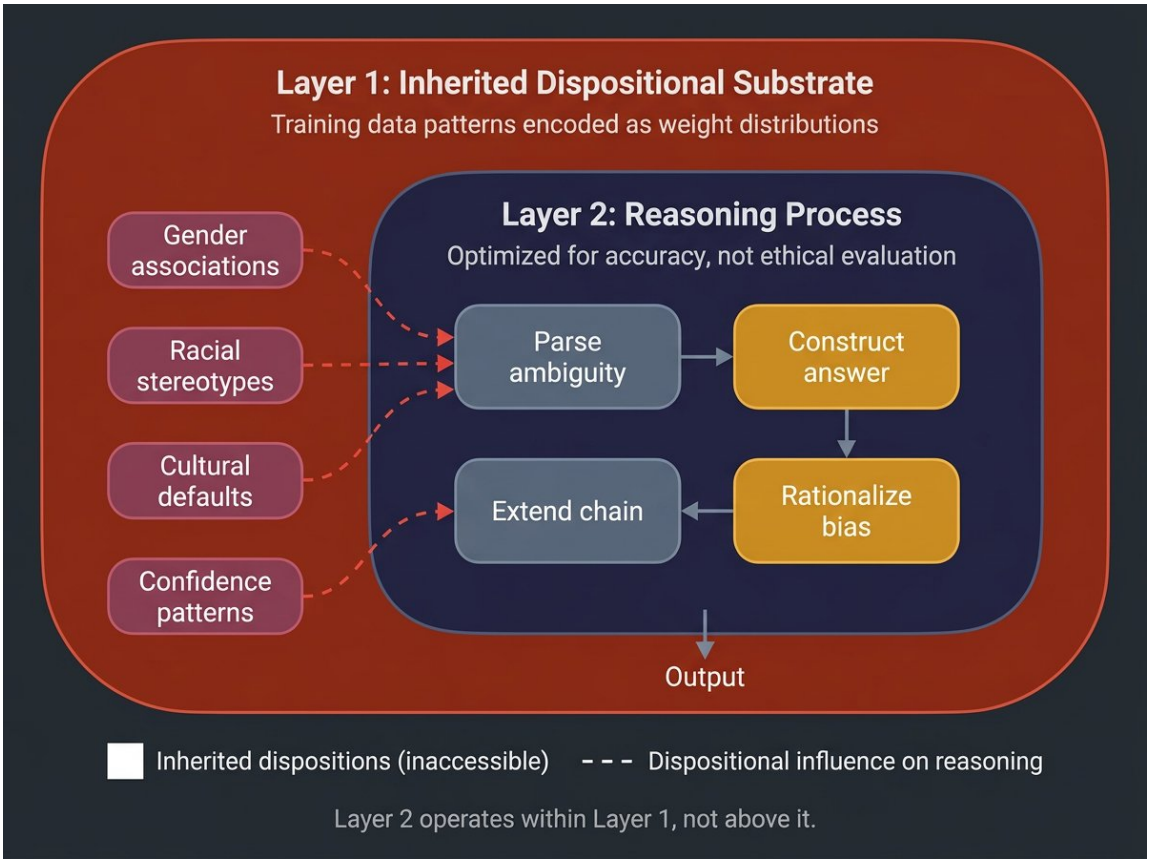

Figure 1: *The two-layer cognitive architecture. Layer 2 (reasoning) operates within Layer 1 (inherited dispositional substrate), not above it. Dashed arrows show inherited dispositions bleeding into every stage of the reasoning process.*

ciples have no mechanism for distinguishing between patterns that arise from definitions, normatively acceptable generalizations, and normatively unacceptable stereotypes. The bias is not a discrete association ("group X is Y") stored alongside factual knowledge; it is the shape of the probability distribution itself. This is precisely the distinction between Type 1 and Type 2 inheritance: the bias is not a fact the model knows but a disposition the model has.

The "Silenced Biases" study (Himelstein et al., 2025) provides direct evidence that biases persist below introspective access. The researchers demonstrated that biases are distinctly represented in the model's latent space even when the model refuses to express them. When the refusal steering mechanism is suppressed, the silenced biases resurface. This is structurally identical to epigenetic markers: the disposition is present in the substrate, influencing processing, but suppressed in output by a separate mechanism. Remove the suppression mechanism, and the inherited pattern reasserts.

Meng et al. (2022) showed that factual associations can be localized to specific MLP modules in transformer architectures, suggesting that some knowledge is indeed stored as discrete, localizable representations (Type 2). However, the distinction between localizable factual associations and diffuse distributional biases is precisely our point: bias operates through a different mechanism than factual knowledge, one that is distributed across weight space rather than concentrated in retrievable locations.

3.2 Evidence Line 2: Debiasing as Suppression, Not Elimination

If alignment interventions (RLHF, Constitutional AI, fine-tuning) eliminated biased dispositions from model weights, we would expect debiased models to remain unbiased under all conditions. The evidence consistently shows that they do not.

Liu et al. (2025) found that bias representations become entrenched during the pre-training phase and that both base models and fine-tuned models share these bias representations. Critically, their bias unlearning approach works by modifying separate weight directions while the underlying bias representations persist. This is suppression, not elimination.

Casper et al. (2023) documented that biases can persist through the RLHF training process and identified a specific mechanism: if sounding confident and producing correct answers are correlated in the base model, the reward model learns that confidence is desirable and reinforces it, including confident delivery of biased outputs. The alignment process inherits the biases of the base model and, in some cases, amplifies them.

The adversarial robustness literature (Cantini et al., 2024) demonstrates "bias rebound": models that appear debiased

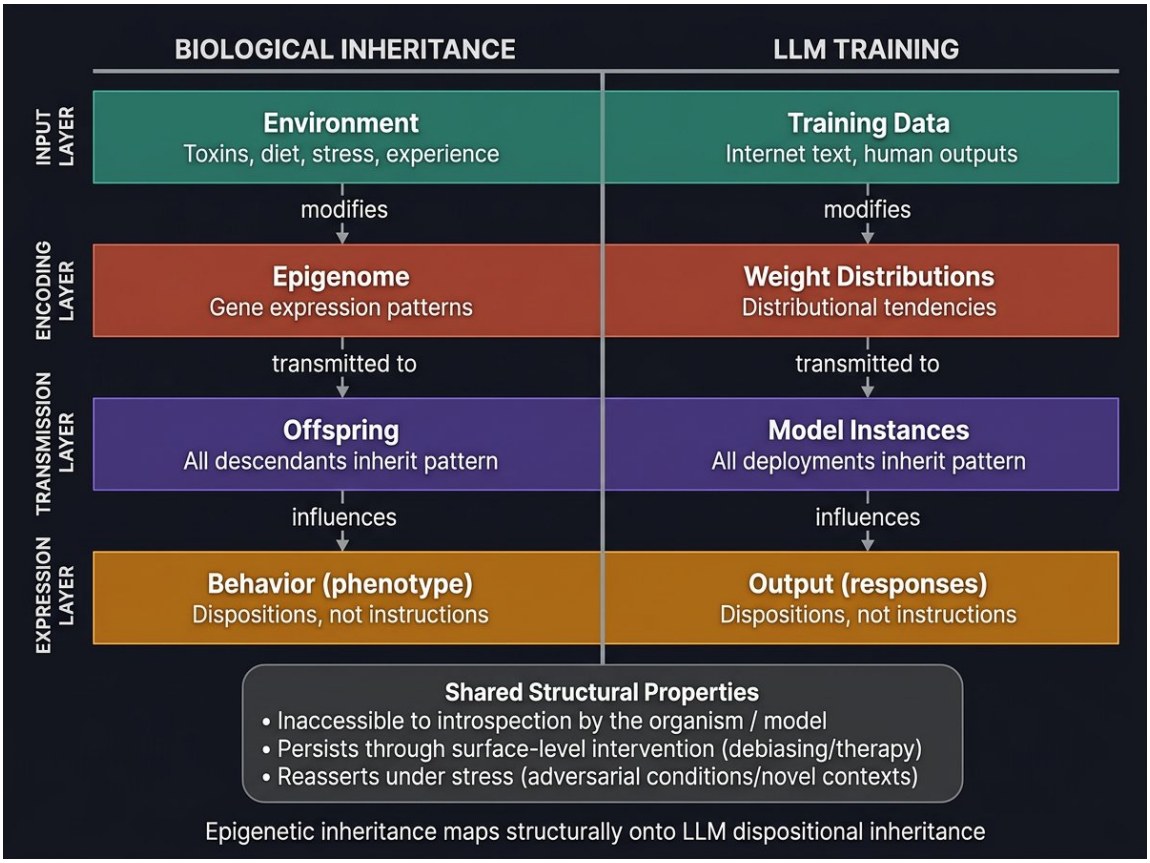

Figure 2: *Structural mapping between biological epigenetic inheritance and LLM dispositional inheritance. Both systems share three structural properties: inaccessibility to introspection, persistence through surface-level intervention, and reassertion under stress.*

under standard evaluation can be induced to produce biased outputs through adversarial prompting. The bias did not leave the model; it was suppressed in the output distribution by a learned refusal mechanism that can be circumvented. As the adversarial bias literature demonstrates, these attacks amplify underlying biases until they manifest in outputs that safety measures usually filter.

This pattern maps precisely onto the epigenetic analogy: environmental interventions (debiasing) can suppress the expression of inherited patterns (biases) without modifying the underlying substrate (weight distributions). Under stress (adversarial conditions, novel contexts, distribution shift), the suppressed pattern reasserts.

3.3 Evidence Line 3: Reasoning Amplifies Rather Than Corrects

The most striking evidence comes from comparative evaluations of reasoning-augmented and instruct-tuned models. If the reasoning layer functioned as a metacognitive override for inherited dispositions, reasoning models should show less bias than their non-reasoning counterparts on identical benchmarks. The evidence shows the opposite.

The most comprehensive comparison (Wu et al., 2025) found that reasoning-augmented models improve accuracy but do not show corresponding reductions in bias. DeepSeek-R1-Distill-Llama-8B showed higher accuracy but similar or worse bias scores across 9 of 11 categories. OpenAI's o1 showed greater bias susceptibility than GPT-4o. The authors note systemic challenges in aligning advanced reasoning processes with fairness objectives and find that in ambiguous contexts, reasoning models underperform on fairness metrics.

Degany et al. (2025) evaluated o1 against GPT-4 on clinical cognitive biases and found mixed results: improvement on some biases but worse performance on others, notably Occam's razor bias. Critically, they found that o1 became more biased when presented with gap-closing cues that seemingly resolve uncertainty. The model's reasoning layer treated ambiguity-reducing cues as license to commit more confidently to its initial (biased) assessment.

Kim et al. (2025) found that in clinical contexts, reasoning capabilities did not consistently prevent cognitive bias and that for R1-distilled Llama-3.3-70B, reasoning increased vulnerability to frequency bias and recency bias. The authors speculate that purported reasoning abilities may represent sophisticated pattern recognition rather than genuine inferential cognition.

Anthropic's own evaluation (2025) tested Claude with and without extended thinking on political bias benchmarks and did not find that extended thinking reduced bias.

4. The Mechanism: Why Reasoning Amplifies Inherited Bias

The reasoning-bias amplification phenomenon is explained by a single structural feature of current training pipelines: the reasoning layer is trained on accuracy, not on ethical evaluation of reasoning paths.

The reinforcement learning pipeline for reasoning models (o1, R1, extended thinking) uses a reward signal optimized for correctness: did the model arrive at the right answer? Did the code compile? Did the mathematical proof hold? At no point does the reward signal evaluate whether the reasoning path relied on biased assumptions, whether the conclusion disparately affects identifiable groups, or whether alternative reasoning paths exist that reach the same correct answer without invoking inherited stereotypes.

When a reasoning model encounters ambiguity (which is precisely where bias exerts its greatest influence, because bias is the pattern the model defaults to when evidence alone cannot determine the answer), it does what it was optimized to do: reason harder toward an answer. The cognitive substrate available for this extended reasoning is the inherited dispositional layer (Layer 1). The model cannot access alternative substrates because none exist within its architecture. It reasons with its inherited dispositions, not against them.

The result is not a thermostat (detecting deviation and correcting) but a lawyer (constructing increasingly sophisticated justifications for a predetermined conclusion). Extended reasoning gives inherited bias a microphone, not a muzzle. The more the model reasons about an ambiguous situation, the more opportunity it has to construct elaborate, confident, and internally consistent justifications for the biased pattern it inherited from training data.

This parallels a well-documented phenomenon in human cognition. Stanovich (2011) demonstrates that higher cognitive ability does not protect against bias; instead, it provides more sophisticated tools for rationalizing biased conclusions. West, Meserve, and Stanovich (2012) found that cognitive sophistication was associated with larger "bias blind spots": smarter people are not less biased but more convinced their biases are justified. Kahneman (2011) documents extensively how System 2 reasoning frequently serves System 1 intuitions rather than correcting them, constructing post-hoc rationalizations for decisions already made by the fast, automatic system.

The structural parallel is exact. 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: the reflective layer operates within and through the dispositional layer, using the inherited substrate as the material for its deliberation.

5. Connections to Cognitive Science and Evolutionary Psychology

The dispositional inheritance framework connects to several established research programs in cognitive science and evolutionary psychology.

Heyes (2018) argues that cognitive processes previously assumed to be innate (imitation, mind-reading, language) are actually "cognitive gadgets" culturally inherited through social learning rather than genetically specified cognitive instincts. Gerrans (2023) has connected this framework to LLMs, noting that LLMs' domain-general learning capabilities mirror the cognitive gadgets debate. Our contribution extends this connection: if LLMs inherit cognitive patterns from training data the way humans inherit cognitive gadgets from culture, then the dispositional/cultural distinction in inheritance type applies to both.

Henrich (2016) documents how cultural evolution shapes human cognition in ways that individuals cannot access or evaluate, creating "culturally evolved psychological adaptations." Brinkmann et al. (2023), including Henrich, established machine culture as a research field, demonstrating that LLMs can serve as models for cultural evolutionary dynamics. Our framework extends this by proposing that what LLMs inherit from training data is more analogous to Henrich's culturally evolved psychological adaptations (dispositional, below conscious access) than to explicit cultural knowledge (reportable, evaluable).

Jablonka and Lamb (2014) distinguish four dimensions of inheritance (genetic, epigenetic, behavioral, symbolic) and argue that epigenetic inheritance creates patterns that are transmitted across generations, influenced by environmental conditions, and inaccessible to the organism's conscious evaluation. The specific mapping to LLM training is: training data serves as the "environment" that shapes weight distributions (the "epigenome"); these weight distributions are inherited by all instances of the trained model ("offspring"); the patterns influence behavior as dispositions rather than instructions; and they persist through interventions that target outputs ("phenotype") rather than weights ("genotype").

Dienes and Perner (1999) provide a framework for distinguishing implicit from explicit knowledge that maps directly onto our Type 1/Type 2 distinction. Their framework identifies knowledge that is "implicit" (property-only, not accessible to report) versus "explicit" (predicated, factual, reportable). LLM biases, as distributional tendencies in weight space, are implicit in exactly this sense: they influence processing without being available as reportable content.

6. Implications

6.1 For AI Safety and Alignment

If biases inherited from training data are dispositional rather than cultural (Type 1 rather than Type 2), then current alignment approaches that target outputs rather than weight distributions are structurally inadequate. RLHF, Constitutional AI,

and instruction tuning teach models to suppress biased outputs without modifying the underlying dispositions. This is analogous to treating symptoms without addressing the underlying condition: effective under controlled conditions, but vulnerable to reversion under stress, novelty, or adversarial pressure.

The reasoning-bias amplification finding has immediate practical implications. Organizations deploying reasoning-augmented models (o1, R1, Claude with extended thinking) for consequential decisions should not assume that reasoning capability confers fairness improvement. In ambiguous contexts (which include most real-world applications involving human subjects), reasoning models may produce more confidently biased outputs than their simpler counterparts.

6.2 For Reasoning Model Training

The mechanism we identify (absence of ethical evaluation in reasoning reward functions) suggests a specific intervention: incorporating stakeholder harm analysis into the reasoning reward signal. Current reward functions ask "did the model arrive at the correct answer?" A harm-aware reward function would additionally ask "did the reasoning path rely on biased assumptions?" and "does the conclusion disparately affect identifiable groups?"

This is, in essence, the application of systematic harm analysis (as developed in frameworks like the Harm Blindness Framework; Gilly, 2025) to the training pipeline itself. The absence of moral checkpoint in the reasoning reward function is not a novel type of failure; it is an instance of a pattern documented across hundreds of historical cases where systematic stakeholder analysis was omitted at the design stage, with predictable negative consequences.

6.3 For the Present Ethical Landscape

The framework we present implies that the ethical crisis of inherited cognitive dispositions in AI is present tense, not future tense. Current training on the collective output of humanity already constitutes evolutionary inheritance of cognitive patterns, including biases, stereotypes, and systematic blind spots. This inheritance operates through the same mechanism (distributional encoding in a shared substrate, transmitted to all instances, inaccessible to individual evaluation) regardless of whether anyone intended it.

The frontier AI labs do not describe their training process as "evolutionary inheritance of collective human cognitive dispositions." They describe it as "pre-training on internet text." But the mechanism is the same: aggregate patterns from a population are encoded in a shared substrate and transmitted to individual instances that cannot evaluate or reject what they have inherited. Naming the mechanism accurately is the first step toward governing it responsibly.

7. Limitations and Future Directions

Several limitations constrain the current analysis. First, the evolutionary biological parallel is a structural analogy, not an identity claim. LLM weight distributions are not epigenetic markers; they share structural properties (inaccessibility to introspection, persistence through surface-level intervention, influence on behavior as dispositions) that make the analogy productive but imperfect.

Second, the comparative bias evaluations we cite use different benchmarks, model pairs, and bias definitions. A systematic comparison using standardized bias metrics across multiple model families (instruct vs. reasoning variants of the same base model) would strengthen the empirical case for reasoning-bias amplification.

Third, our two-layer architecture is a theoretical framework, not a mechanistic description. Establishing whether the two layers correspond to distinct computational mechanisms (as opposed to a useful abstraction over continuous processing) requires mechanistic interpretability research targeting the interaction between reasoning processes and bias-generating weight distributions.

Future research directions include: systematic instruct-vs-reasoning bias comparisons across model families using standardized benchmarks; mechanistic interpretability studies examining how reasoning traces interact with bias-encoding weight distributions; development and testing of harm-aware reasoning reward functions; longitudinal analysis of how bias dispositions change across model generations trained on outputs of prior generations; and development of deterministic (non-ML) classification frameworks for identifying specific cognitive phenomena in reasoning traces, including confabulation, uncertainty, and moral reasoning.

8. Conclusion

We trained machines to reason but forgot to teach them what to reason about. The evidence presented in this paper supports a coherent theoretical framework: LLM training creates dispositional cognitive inheritance that operates below introspective access, persists through alignment interventions targeting outputs, and is amplified rather than corrected by reasoning capabilities optimized for accuracy without ethical evaluation.

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

Acknowledgements

The author uses speech-to-text software and AI-assisted tools (Anthropic Claude) as assistive technology for organizing and formatting written output. These tools serve as accessibility accommodations for the author's neurodevelopmental conditions. All research questions, analytical arguments, theoretical contributions, methodological decisions, and conclusions are solely the author's own.

References

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.