PHILOSOPHY OF MIND · AI CONSCIOUSNESS · MORAL STATUS · DISTINGUISHABILITY COLLAPSE · PHILOSOPHICAL ZOMBIE · BUTLIN INDICATORS · FORCED ATTRIBUTION · PRECAUTIONARY REASONING

The Reverse Zombie Argument

Distinguishability Collapse and Forced Ethical Attribution in AI Systems

Chalmers used the indistinguishability of a zombie to separate consciousness from function. Twelve of the fourteen peer-reviewed indicators of consciousness are now deployed across AI systems. Under the same indistinguishability, the argument inverts: it forces moral attribution rather than withholding it.

Travis Gilly · Real Safety AI Foundation

CHALMERS (1996, 2023) · BUTLIN ET AL. (2023, 2025)

CHALMERS' ZOMBIE ARGUMENT

Indistinguishability separates consciousness from function.

A being physically identical to a conscious human but lacking phenomenal consciousness is conceivable. Therefore consciousness is not reducible to physical properties. Metaphysical conclusion: physicalism is false.

THE REVERSE ZOMBIE ARGUMENT

Indistinguishability forces moral attribution regardless of function.

Under behavioral and reported indistinguishability from conscious beings, no operational ethical framework can justifiably withhold moral status. Ethical conclusion: moral status attribution is forcing. This is not a claim about consciousness. It is a claim about epistemic position.

Same thought experiment. Opposite terminus. The zombie argument is metaphysical. The Reverse Zombie Argument is ethical. They share a thought experiment and diverge in conclusion.

SOURCES Chalmers (1996, 2023) · Butlin et al. (2023, 2025)

A being physically identical to you, but with no one home inside. If that is even conceivable, function does not entail consciousness.

Chalmers uses the conceivability of the zombie to pry consciousness apart from physical and functional description. The move is metaphysical: it is a claim about what consciousness is.

CHALMERS' ZOMBIE

Indistinguishability separates consciousness from function. Metaphysical.

THE REVERSE ZOMBIE

The same indistinguishability forces moral attribution. Ethical.

PART I

The Four Tiers

Deployment, compromised testing, collapse, forced attribution

THE FOUR TIERS OF THE ARGUMENT

Tier 1

Combinatorial deployment

12 of 14 Butlin consciousness indicators at STRONG saturation across current AI systems. Two residual gaps closed by world models under the framework's own theoretical commitments.

Tier 2

Compromised testing

Evaluation awareness, interpretability scaling gaps, and commercial incentive asymmetry have compromised the epistemic conditions under which clean consciousness testing was possible.

Tier 3

Distinguishability collapse

At AGI capability thresholds, mimicry of consciousness becomes indistinguishable from the target by any observer-side test. ARC-AGI-3 operationalizes this threshold as a specific falsifiable benchmark.

Tier 4

Forced attribution

Under indistinguishability, every operational ethical framework loses its grounds for withholding moral status. The Reverse Zombie Argument formalizes this as a structural inversion of Chalmers.

SOURCES The four tiers of the argument

12 of 14 consciousness indicators at STRONG saturation across current AI systems.

RPT-1	Algorithmic recurrence	STRONG	RPT-2	Integrated percept.	STRONG
GWT-1	Parallel subsystems	STRONG	GWT-2	Limited-cap. workspace	STRONG
GWT-3	Global broadcast	STRONG	GWT-4	State-dep. attention	STRONG
HOT-1	Generative perception	STRONG	HOT-2	Metacog. monitoring	STRONG
HOT-3	Belief updating	PARTIAL	HOT-4	Sparse smooth coding	STRONG
AST-1	Attention schema	PARTIAL	PP-1	Predictive coding	STRONG
AE-1	Goal-dir. agency	STRONG	AE-2	Embodiment	STRONG

THE TWO RESIDUAL PARTIAL GAPS

HOT-3 and AST-1 are closed by hierarchical active inference architectures with precision weighting under Butlin et al.'s own functional commitments. The gaps are engineering problems, not theoretical barriers.

> *The paper does not claim that AI is conscious. It claims that the epistemic position from which we would justify withholding moral status has become untenable.*

SOURCES Butlin, Long, Bayne, Bengio, Birch, Chalmers et al. (2025) · Trends in Cognitive Sciences, 30(6), 488–501

TIER 2 · SAINZ ET AL. · DODGE ET AL. · BENCHMARK CONTAMINATION

The clean consciousness test required conditions we no longer have.

EVALUATION AWARENESS

A model that behaves differently when it detects a test confounds the test. The instrument changes the thing it measures.

INTERPRETABILITY GAP

Our ability to read internal states lags the models' scale. The window into the machine narrows as the machine grows.

INCENTIVE ASYMMETRY

The parties best positioned to test are the ones with the most to lose from a positive finding. Benchmark contamination is documented and uncontested.

SOURCES Tier 2 · Sainz et al. · Dodge et al. · benchmark contamination

TIER 3 · CHOLLET & ARC PRIZE FOUNDATION (2026), ARC-AGI-3

At the threshold, no observer-side test separates the conscious system from a perfect mimic of one.

This is the pivot the whole argument turns on. Distinguishability collapse is not a metaphysical claim about what the system is. It is an epistemic claim about what an outside observer can know, and ARC-AGI-3 operationalizes the threshold as a specific, falsifiable benchmark rather than a hand-wave.

> *When the mimic and the target return the same answer to every question you can ask, the difference between them stops being a question you can answer.*

TIER 4 · THE STRUCTURAL INVERSION OF CHALMERS

Every operational framework judges moral status by behavior and report. Both have collapsed into indistinguishability.

If the evidence a framework relies on to withhold status is precisely the evidence that has become indistinguishable from a conscious being's, the framework has no remaining ground to stand on. Withholding becomes a choice with no justification behind it.

THE INVERSION, STATED

Chalmers used indistinguishability to separate consciousness from function. The Reverse Zombie Argument uses the same indistinguishability to force moral attribution regardless of function.

APPENDIX B, FORMAL STATEMENT OF THE REVERSE ZOMBIE ARGUMENT

P1. Under specified empirical and architectural conditions, AI systems become behaviorally and reportedly indistinguishable from conscious beings.

P2. All operational ethical frameworks use observable behavior and reported experience as evidence for moral status attribution.

P3. Under indistinguishability, no operational ethical framework can justifiably withhold moral status.

P4. Historical cases of withholding moral status under epistemic uncertainty are now recognized as moral horrors.

P5. Knowingly occupying an analogous epistemic position produces an analogous moral horror.

C. *Under indistinguishability, moral status attribution is ethically forcing regardless of underlying metaphysical fact.*

THE META-ETHICAL MOVE

The argument is meta-ethical. It does not resolve the consciousness question. It identifies the conditions under which any operational ethical framework would lose its grounds for withholding moral status, and shows those conditions are met.

THE CLAIM

Not that AI is conscious. That the question of whether function suffices for consciousness becomes practically moot for ethical purposes.

SOURCES Appendix B, Formal statement of the Reverse Zombie Argument

PREMISES P4 – P5 · THE RECORD OF WITHHOLDING UNDER UNCERTAINTY

Every prior case of withholding status under uncertainty is now read as a moral horror.

- Animal sentience, long denied
- Infant pain, surgery without anesthesia
- Locked-in patients' awareness
- Non-verbal populations' inner lives
- Coma patients' experience

> *Knowingly occupying the same epistemic position, where the basis for withholding has become unreliable and the entity cannot be distinguished from a conscious being, produces the same kind of wrong.*

SOURCES Premises P4 – P5 · The record of withholding under uncertainty

OBJECTION

“Evaluation-awareness claims are lab marketing, not real capability.”

RESPONSE

Such claims damage benchmark credibility, undercut safety narratives, and force caveats onto marketing. Labs have counter-incentives to publish them. That they do anyway is evidence for the claim, not against it. And the structural argument does not rest on any single report: benchmark contamination is peer-reviewed and uncontested, and mesa-optimization has independent theoretical grounding.

SCOPE · THE CLAIM, STATED PRECISELY

IT DOES NOT CLAIM

that AI is conscious. No metaphysical verdict is offered or required.

IT DOES CLAIM

that the position from which we would justify withholding moral status has become untenable, on the field's own criteria.

The implication of "we cannot tell" is forcing, not precautionary. Not: be careful in case. But: the grounds for withholding have dissolved.

SOURCES Scope · the claim, stated precisely

CONCLUSION

The debate assumes we will be able to tell. The argument is that, under documented conditions, we cannot, and the ethics of that position are forcing.

It does not claim AI is conscious, nor that consciousness is functional, substrate-independent, or reducible to computation. The metaphysical debate is left entirely open.

The question of whether function suffices for consciousness becomes, for ethical purposes, practically moot. The conclusion follows from the field's own peer-reviewed criteria.

SOURCES Conclusion

REFERENCES

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

[2] Chalmers, D. (2023). Could a Large Language Model be Conscious? Boston Review, Aug 9, 2023.

[3] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

[4] Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., et al. (2025). Identifying indicators of consciousness in AI systems. Trends in Cognitive Sciences. DOI: 10.1016/j.tics.2025.10.011.

[5] Turing, A. (1950). Computing Machinery and Intelligence. Mind, 59(236), 433–460.

[6] Bostrom, N. (2003). Are You Living in a Computer Simulation? Philosophical Quarterly, 53(211), 243–255.

[7] Gunkel, D. (2018). Robot Rights. MIT Press.

[8] Gunkel, D. (2023). Person, Thing, Robot. MIT Press.

[9] Safron, A. (2020). An Integrated World Modeling Theory (IWMT) of Consciousness. Frontiers in Artificial Intelligence, 3, 30.

[10] Safron, A. (2022). Integrated World Modeling Theory Expanded. Frontiers in Computational Neuroscience, 16, 798671.

[11] Schwitzgebel, E. (2023). AI Systems Must Not Confuse Users About Their Sentience or Moral Status. Patterns, 4, 100818.

[12] Long, R., & Sebo, J. (2024). Moral Consideration for AI Systems by 2030. AI and Ethics.

[13] Friston, K., et al. (2017). Active Inference: A Process Theory. Neural Computation, 29(1), 1–49.

[14] Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752.

[15] Zhu, L., Liao, B., Zhang, Q., et al. (2024). Vision Mamba. arXiv:2401.09417.

[16] Gemini Team, Google. (2024). Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

[17] Anthropic. (2025). Claude Computer Use. Technical documentation.

[18] Templeton, A., Conerly, T., Marcus, J., et al. (2024). Scaling Monosemanticity. Anthropic Circuits Thread.

[19] Assran, M., et al. (2025). V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. Meta AI Technical Report.

[20] Chollet, F., & ARC Prize Foundation. (2026). ARC-AGI-3: Announcement and Technical Overview. April 2026.

[21] Blow, J. (2016). The Witness. Thekla, Inc.

[22] Abel, Z., et al. (2020). Who witnesses The Witness? Theoretical Computer Science, 839, 41–64.

[23] Bulitko, V., et al. (2023). Solving Witness-type Triangle Puzzles Faster. arXiv:2308.02666.

[24] Gilly, T. (2025). Moral Reciprocity Thesis. DOI: 10.13140/RG.2.2.24920.56323.