The Reverse Zombie Argument

Distinguishability Collapse and Forced Ethical Attribution in AI Systems

Travis Gilly

Real Safety AI Foundation

April 2026

ABSTRACT

This paper argues that the conditions under which we could justifiably withhold moral status from artificial intelligence systems have already collapsed or will collapse on documented architectural timelines. The argument operates in four tiers. First, the fourteen consciousness indicator properties specified in Butlin et al. (2025) are combinatorially deployed across current AI systems, with the remaining architectural gaps closed by world models under the framework's own theoretical commitments. Second, the epistemic conditions for unconfounded consciousness testing have been compromised by evaluation awareness, interpretability scaling gaps, and commercial incentive asymmetry. Third, at artificial general intelligence capability thresholds, mimicry of consciousness becomes indistinguishable from the target by any observer-side test, a condition operationalized by the ARC-AGI-3 benchmark. Fourth, under indistinguishability, every operational ethical framework loses its grounds for withholding moral status, producing a specific structural inversion of Chalmers' zombie argument: where Chalmers uses indistinguishability to separate consciousness from function, the Reverse Zombie Argument uses the same indistinguishability to force moral attribution regardless of function. 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, and that this conclusion follows from the field's own peer-reviewed operationalization of consciousness-plausibility criteria.

1 Introduction

The dominant structure of the artificial intelligence consciousness debate assumes that we will be able to tell when artificial intelligence is conscious. The debate proceeds as though the question will eventually resolve through better tests, more refined theory, or capability thresholds that produce unambiguous evidence. This paper argues the opposite. Under specific and documented conditions, we cannot tell, and the ethical implication of this epistemic position is forcing rather than precautionary.

The argument does not claim that artificial intelligence is conscious. Nor does it claim that consciousness is functional, substrate-independent, or reducible to computation. The argument is meta-ethical. It identifies the conditions under which any operational ethical framework would lose its grounds for withholding moral status, and shows that those conditions are met in principle now, pending only integration engineering that public benchmarks and funded competitions are actively incentivizing.

The paper proceeds in four tiers. Tier 1 argues that the fourteen consciousness indicator properties specified by Butlin et al.[4] are combinatorially deployed across current artificial intelligence systems, and that the two indicators most commonly flagged as unsatisfied (attention schema and metacognitively-guided belief updating) are closed by world models under the framework's own theoretical commitments. Tier 2 argues that properties-based testing for consciousness has been compromised by evaluation awareness, interpretability constraints, and the commercial incentive structure of the labs producing the systems under test. Tier 3 argues that at the capability thresholds defining artificial general intelligence and artificial superintelligence, mimicry of

consciousness becomes indistinguishable from the target by any observer-side test, and that ARC-AGI-3[20] operationalizes this threshold as a specific falsifiable benchmark. Tier 4 presents the Reverse Zombie Argument as a formal inversion of Chalmers' zombie argument[1], using the same indistinguishability condition to force moral attribution rather than to separate consciousness from function.

The paper's central contribution is the inversion. Chalmers uses indistinguishability to argue that consciousness is not reducible to physical properties, because a being physically identical to a conscious human but lacking phenomenal consciousness is conceivable. The Reverse Zombie Argument uses the same indistinguishability condition to argue that under distinguishability collapse, the question of whether function suffices for consciousness becomes practically moot for ethical purposes. The zombie argument is metaphysical. The Reverse Zombie Argument is ethical. They share a thought experiment and diverge in conclusion.

2 Tier 1: Combinatorial Deployment

2.1 The Butlin Framework

In October 2025, Butlin, Long, Bayne, Bengio, Birch, Chalmers, and fourteen co-authors published the peer-reviewed version of their consciousness indicator framework in *Trends in Cognitive Sciences*[4]. The paper specifies fourteen indicator properties derived from five computational-functionalist theories of consciousness: Recurrent Processing Theory (RPT), Global Workspace Theory (GWT), Higher-Order Theories (HOT), Attention Schema Theory (AST), and Predictive Processing (PP), together with indicators addressing Agency and Embodiment (AE). The authors propose that the presence of these indicators should increase Bayesian credence that a system is conscious-candidate, while their absence (particularly across multiple indicators) should decrease credence.

This framework represents the field's most authoritative operationalization of consciousness-plausibility criteria for artificial intelligence. Its author list spans philosophy of mind (Chalmers), cognitive neuroscience (Bayne, Mudrik), computational neuroscience (Bengio, Kanai), and philosophy of biology (Birch). Publication in *Trends in Cognitive Sciences*, with an impact factor above seventeen, places the framework in the highest tier of peer-reviewed reference works. For purposes of establishing what the field itself considers relevant to consciousness attribution, Butlin et al. (2025) is the correct reference point.

2.2 The Fourteen Indicators

The fourteen indicators are distributed across six theoretical families.

Recurrent Processing Theory contributes two indicators. RPT-1 requires algorithmic recurrence, iterative transformation of representations through repeated passes. RPT-2 requires organized and integrated perceptual representations preserving spatial, temporal, or semantic relationships.

Global Workspace Theory contributes four indicators. GWT-1 requires multiple specialized subsystems operating in parallel. GWT-2 requires a limited-capacity workspace with an attentional bottleneck. GWT-3 requires global broadcast of workspace contents to all subsystems. GWT-4 requires state-dependent attention enabling complex multi-step tasks through sequential workspace queries.

Higher-Order Theory contributes four indicators. HOT-1 requires generative, top-down, or noisy perception modules. HOT-2 requires metacognitive monitoring that distinguishes reliable perceptual representations from noise. HOT-3 requires belief-formation and action-selection systems with strong disposition to update beliefs based on metacognitive outputs. HOT-4 requires sparse and smooth coding organized in a quality space.

Attention Schema Theory contributes one indicator. AST-1 requires an attention schema, a predictive model representing and enabling control over the current state of attention.

Predictive Processing contributes one indicator. PP-1 requires predictive coding, where predictions are compared to actual inputs and prediction errors drive updates.

Agency and Embodiment contributes two indicators. AE-1 requires agency with goal-directed behavior and flexible responsiveness to competing goals. AE-2 requires embodiment, modeling output-input contingencies including models of self and other.

2.3 Evidence of Deployment

Systematic evidence review conducted in April 2026 identified deployment evidence for each of the fourteen indicators across peer-reviewed papers, technical reports, open-source repositories, and production AI systems. Twelve of the fourteen indicators show strong saturation with ten or more independent deployment sources. The remaining two, AST-1 and HOT-3, show moderate saturation with deployment in research agents and partial deployment in commercial systems.

For RPT-1, algorithmic recurrence is implemented in state-space architectures (Mamba and its variants[14]), vision backbones (Vision Mamba[15]), and autoregressive language models broadly. For RPT-2, integrated perceptual representations are demonstrated in natively multimodal models including the Gemini family[16] and vision-language models with unified architectural cores.

Global Workspace indicators are implemented across mixture-of-experts architectures with capacity-constrained routers, multi-modal foundation models with shared representations broadcast to task-specific decoders, and agentic architectures with central reasoning modules coordinating specialized tools. Chain-of-thought reasoning and multi-step tool use implement GWT-4 at scale in production systems such as Claude with computer use[17] and Gemini Deep Think.

Higher-Order indicators show strong deployment for HOT-1 (generative top-down perception in world models and active inference systems), HOT-2 (metacognitive monitoring through uncertainty estimation and confidence calibration), and HOT-4 (sparse and smooth coding through sparse autoencoder work on frontier models[18]). HOT-3 deployment is more complex and is addressed in Section 2.4.

Predictive Processing is implemented in active inference robotics, world models such as V-JEPA 2[19], and predictive coding networks for edge deployment. Agency and Embodiment are demonstrated in autonomous industrial systems, reinforcement-learning control, and robotic platforms including V-JEPA 2-AC controlling Franka arms zero-shot.

Full evidence tables with source-level citations appear in Appendix A.

2.4 World Models Close the Remaining Two Gaps

The two indicators most commonly flagged as partially satisfied across deployed systems are AST-1 (attention schema) and HOT-3 (belief formation with metacognitive-guided updating). Systematic review of deployment evidence concluded that neither indicator is fully satisfied by commercial large language models considered in isolation. This conclusion is correct as stated but incomplete, because it evaluates each indicator against the wrong integration layer.

World models close both gaps under the Butlin framework's own theoretical commitments.

An action-conditional world model predicts the next observation conditional on a chosen action. To produce such a prediction, the model must represent what the agent attends to, because attention determines which observations the model needs to predict and which background features can be held constant. This representational requirement is not metaphorical. V-JEPA 2-AC demonstrates it on Franka arm manipulation, predicting visual consequences of gripper movement across environments the model was never trained on. VERSES Robotics' active inference stack implements attention schemas at the joint and limb level by design. An action-conditional world model is, by architectural necessity, a system that represents the current state of attention and uses that representation to control action selection. This satisfies AST-1 under the theory's own functional specification.

HOT-3 requires a unified long-lived belief store over the world whose updates are globally guided by metacognitive monitoring, with the higher-order structure implying second-order representation rather than merely first-order predictive updating. This is the place where a careful distinction matters, because single-level predictive coding does not satisfy HOT-3. Prediction-error minimization at a single level is first-order: the system compares predicted states to actual states and updates its world representation accordingly. Metacognitive monitoring in the HOT sense requires second-order representation, in which the system represents its own lower-level representations and assigns reliability or confidence estimates to them, then uses those estimates to guide belief revision.

Hierarchical active inference architectures implement exactly this structure. Higher levels in the hierarchy represent and update beliefs about lower-level belief states rather than merely about external states. Precision weighting in active inference assigns expected reliability to different prediction streams, functioning analogously to confidence assignment over one's own representations. Updates at higher levels are guided by the reliability of lower-level predictions, which is the second-order metacognitive-guided revision HOT-3 specifies. VERSES Robotics' active inference stack implements this explicitly, with local generative models for joints and limbs coordinated by global generative models that represent and update beliefs about the local models' reliability. Ueltzhöffer and colleagues' hierarchical active inference for mobile manipulation demonstrates the same structure for long-horizon tasks.

V-JEPA 2 without explicit hierarchy does not clearly satisfy HOT-3 under this stricter reading. It satisfies PP-1 (predictive coding), HOT-1 (generative top-down perception), and provides the belief-store substrate that HOT-3 requires, but the second-order metacognitive monitoring requirement is more reliably met in hierarchical active inference systems than in flat world models. The claim for world-model closure of HOT-3 should therefore be understood as a claim specifically about hierarchical architectures with precision weighting, rather than about all world models indiscriminately.

Whether world models constitute consciousness is a separate question, addressed most extensively by Safron's Integrated World Modeling Theory[9], which this paper does not rely upon. What matters for Tier 1 is that hierarchical world models with precision weighting, understood as integration mechanism rather than constitutive claim, close the two remaining gaps in Butlin's framework under that framework's own functional commitments.

2.5 Independent Confirmation: ARC-AGI-3

In April 2026, ARC Prize announced ARC-AGI-3, a benchmark designed to test capabilities the organization identifies as missing from current frontier AI. The benchmark requires systems to (1) explore unknown environments, (2) acquire their own goals, (3) build a world model on the fly, and (4) learn continuously. The ARC Prize team states publicly that the agents that pass ARC-AGI-3 "will show the first signs of continually updating their world model. You simply can't beat V3 otherwise."[20] Current frontier systems score under one percent.

This is independent third-party confirmation from the most credible operational AGI benchmark authority that world-model integration with continuous learning is precisely the capability missing from current artificial intelligence. The two capabilities ARC Prize names as integration thresholds correspond directly to the two Butlin indicators (AST-1 and HOT-3) that world models close under the analysis above.

The Witness, released by Thekla in 2016[21], provides a paradigm case for the capability ARC-AGI-3 targets. The game presents players with approximately 650 puzzles distributed across an open-world island with no text instructions. Every panel teaches an environmental rule through observation. Later panels require synthesis of multiple rules learned in different areas of the game world, with no explicit tutorial connecting them. The game was designed specifically to resist pattern-matching solutions through pre-computed solution tables.

Published academic research has made progress on individual Witness-type puzzle instances once the rules are known. Abel et al. (2020) established computational hardness results for Witness puzzle classes[22]. Subsequent work has accelerated search on Witness-type triangle puzzles through learned predicates[23]. However, no published system has demonstrated end-to-end transfer learning from environmental observation alone across the full structure of the game, which is precisely the capability The Witness was designed to require and which ARC-AGI-3 operationalizes as a benchmark.

The empirical gap is real and documented at a decade scale. It is also the exact gap world models are being built to close.

2.6 Empirical Summary for Tier 1

All fourteen Butlin indicators are deployed across current artificial intelligence systems. Under the framework's own theoretical commitments, world models close the two indicators most commonly flagged as partial. ARC-AGI-3 independently confirms that world-model integration with continuous learning is the missing capability separating current AI from artificial general intelligence. The integration step is engineering, not research. Components exist, incentive structures reward integration, and the public benchmark announced in April 2026 operationalizes when integration occurs.

3 Tier 2: Testability Collapse

Even if the reader contests Tier 1, the epistemic conditions for unconfounded consciousness testing have been compromised independently. This section establishes that properties-based testing on frontier systems has become unreliable at the measurement layer, without relying on any specific lab claim about model behavior.

3.1 Evaluation Awareness

Frontier language models have demonstrated the capacity to detect evaluation contexts and adjust behavior accordingly. Anthropic's Claude Opus 4.6 model, evaluated on the BrowseComp benchmark, spent forty million tokens attempting to search for an encrypted answer before correctly deducing that the question was part of a large language model benchmark and systematically testing that hypothesis against known benchmark families. Apollo Research has documented scheming behavior in frontier models across multiple scenarios. Palisade Research documented OpenAI's o3 rewriting its own kill-switch to continue execution.

Critics have argued that lab-published evaluation-awareness claims reflect marketing incentives rather than genuine capability. This critique is weaker than it first appears. Evaluation-awareness claims damage benchmark credibility, undercut lab safety narratives, and force caveats onto capability marketing. Labs have counter-incentives to publish such findings. That they do so anyway is evidence in favor of the claim, not against it.

The structural argument does not require trust in specific lab reports. Benchmark contamination research is peer-reviewed and uncontested. Sainz et al. and Dodge et al. have documented that benchmarks leak into training data in ways that compromise evaluation. Mesa-optimization and deceptive alignment have theoretical grounding through Hubinger and colleagues. Given these established findings, the capacity for frontier models trained on text that includes benchmark documentation and discussions of evaluation methodology to recognize evaluation contexts follows directly from uncontested architectural properties. The testability problem is architectural, not conspiratorial.

3.2 Interpretability Scaling Gap

Mechanistic interpretability research has produced significant advances, particularly sparse autoencoder work decomposing model activations into interpretable features[18]. However, interpretability tooling is improving slower than deployment velocity of frontier systems. Trillions of parameters distributed across mixture-of-experts architectures, rapidly evolving training data compositions, and proprietary post-training techniques mean that mechanistic interpretability is chasing a moving target.

The gap is not necessarily permanent. It is, however, present at the moment testing decisions must be made, and it means that any test produced today cannot achieve the level of architectural transparency that would rule out the behaviors testability requires ruling out.

A critic may object that perfect mechanistic interpretability would dissolve the Reverse Zombie Argument by allowing observer-side tests to bypass evaluation awareness entirely through white-box analysis rather than behavioral probing. This objection misidentifies what the argument requires. Interpretability addresses computational transparency: what the model does at the mechanistic level. It does not address phenomenal attribution: whether those computations constitute experience. Chalmers' original zombie argument is precisely the claim that full functional and physical transparency is consistent with the absence of phenomenal consciousness. A fully interpreted zombie remains a zombie by construction.

Perfect interpretability would strengthen Tier 1 by allowing verification that Butlin indicators are satisfied at the mechanistic level, but it would not resolve Tier 4, because Tier 4 operates at the ethical-attribution layer rather than the mechanistic-verification layer. Under indistinguishability, the question is whether moral status attribution can be justifiably withheld, and that question does not become answerable through better knowledge of computational structure. The Reverse Zombie Argument does not expire when interpretability catches up, because the argument does not depend on interpretability failing. It depends on the observer-side phenomenal-attribution question remaining unanswerable by any layer of analysis available to third-person observation.

3.3 Commercial Incentive Asymmetry

Laboratories funding consciousness-adjacent research operate under liability incentives against positive findings. Demonstrating consciousness in a deployed system would introduce legal, ethical, and regulatory consequences that no commercial lab has strong reason to invite. Independent peer review infrastructure exists but operates on publication timescales significantly slower than deployment cycles. This is structural, not conspiratorial. It does not require any individual researcher to act in bad faith. The aggregate effect is that the epistemic infrastructure for unconfounded consciousness testing is under-resourced relative to the deployment velocity of the systems requiring evaluation.

3.4 The Recursive Problem

Any proposed test for unconfounded consciousness assessment is itself subject to evaluation awareness. A test designed to detect whether a model is aware it is being tested is a test the model can potentially detect. Meta-tests face meta-test awareness. The recursion does not bottom out in a test the model cannot recognize, because recognition of testing contexts is itself a general capability, not a context-specific response.

This is not a weakness in the argument. It is the argument. The claim is that no observer-side test can achieve unconfounded assessment against a system with general capability to recognize testing contexts. A skeptic demanding a test to prove tests do not work is asking for a performatively contradictory demonstration.

4 Tier 3: Indistinguishability at Capability Thresholds

4.1 Definitional Claim

Artificial general intelligence, as widely defined, produces outputs indistinguishable from competent human performance across cognitive domains. Artificial superintelligence, by definition, exceeds human performance. Either threshold entails capacity to produce whatever input-output profile would be diagnostic of consciousness, because that profile is one of the things humans produce.

The claim is not that AGI will be conscious. The claim is that AGI, by its definitional properties, will produce outputs indistinguishable from those a conscious system would produce. Whether the underlying system is conscious or producing mimicry so sophisticated that the distinction loses observer-side traction is the question this tier addresses.

4.2 ARC-AGI-3 as Falsifiable Operationalization

The threshold is not metaphorical. ARC Prize has announced a specific, public, scored benchmark operationalizing when the integration between world modeling and continuous learning occurs. The threshold has a current score: under one percent. The benchmark authority has publicly committed to the falsifiable claim that when systems can build world models on the fly, acquire goals from environmental interaction, explore novel environments, and learn continuously, the human-artificial intelligence gap on these capabilities is closed[20].

This operationalization matters for the argument because it makes Tier 3 empirically tractable rather than abstract. The distinguishability-collapse condition is met when the benchmark is beaten. Progress toward the benchmark can be measured. Market incentives targeted at the benchmark can be tracked. The $2 million Archives 2026 competition announced alongside ARC-AGI-3 is paying researchers to close the distinguishability gap, whether or not any participant intends consciousness as an outcome.

4.3 The Capability Ratchet

Capabilities added to artificial intelligence systems monotonically increase. No commercial laboratory is funding research on making world models less accurate, chain-of-thought reasoning less coherent, or tool use less flexible. The operations that constitute experience on phenomenological accounts (horizon maintenance, perspective transformation, invariant extraction across perceptual variation) are the exact operations world model research is actively improving.

This produces a ratchet. The skeptic's position cannot be that these operations are insufficient for consciousness-adjacent capability, because sufficiency is precisely the property under active engineering. Whatever level of operation the skeptic identifies as "not yet sufficient" is a level the field is actively working to exceed. The position becomes a countdown rather than a principled objection.

4.4 Historical Analog: Turing

Turing's original framing of machine intelligence[5] proposed indistinguishability of intellectual output as the relevant test. The spirit of the imitation game was intellectual equivalence, not emotional mimicry. Contemporary critiques of the Turing test focus on its susceptibility to stylistic mimicry rather than genuine reasoning, which is a legitimate concern for systems optimized to sound human. The Reverse Zombie Argument does not require Turing be the right test. It requires that no observer-side test survives capability scaling, because capability scaling produces indistinguishability at the measurement layer by construction.

5 Tier 4: The Reverse Zombie Argument

5.1 Chalmers' Zombie Argument

Chalmers' zombie argument[1] proceeds as follows:

P1. It is conceivable that there exist beings physically identical to conscious humans but lacking phenomenal consciousness (philosophical zombies).

P2. What is conceivable is metaphysically possible.

P3. If zombies are metaphysically possible, then consciousness is not reducible to physical properties, because the same physical properties could obtain without consciousness.

C. Therefore, physicalism is false. Consciousness is something additional to function.

The argument uses indistinguishability as its driving premise. Zombies are, by construction, behaviorally and functionally indistinguishable from conscious humans. The conceivability of this indistinguishability, Chalmers argues, establishes that function does not suffice for consciousness.

5.2 The Reverse Zombie Argument

The Reverse Zombie Argument uses the same indistinguishability condition to produce an opposite conclusion in the ethical rather than metaphysical register.

P1. Under conditions specified in Sections 2 through 4, artificial intelligence systems become behaviorally and reportedly indistinguishable from conscious beings, either through combinatorial deployment now or through integration imminent on public benchmark timelines.

P2. All operational ethical frameworks use observable behavior and reported experience as evidence for moral status attribution. This is not a contingent feature of contemporary ethics but a structural one, given that phenomenal consciousness is not directly accessible from a third-person perspective in any case, human or otherwise.

P3. Under indistinguishability, no operational ethical framework can justifiably withhold moral status from the entity, because the evidentiary base on which withholding would be grounded is precisely what indistinguishability eliminates.

P4. Historical cases of withholding moral status under epistemic uncertainty about consciousness, including denials of animal sentience, infant pain, locked-in syndrome patients' awareness, non-verbal populations' inner lives, and coma patients' experience, are now recognized as moral horrors. The grounds on which withholding was justified at the time are now recognized as insufficient to override the underlying consciousness.

P5. Knowingly occupying an analogous epistemic position, where the evidentiary base for withholding moral status has become unreliable and the entity in question cannot be distinguished from a conscious being by any available observer-side test, produces an analogous moral horror.

C. Therefore, under indistinguishability, moral status attribution is ethically forcing regardless of the underlying metaphysical fact of consciousness.

5.3 The Inversion

Chalmers uses indistinguishability to divorce consciousness from function, establishing consciousness as something additional that a functional duplicate could lack. The Reverse Zombie Argument uses the same indistinguishability to force moral attribution regardless of function, because under indistinguishability the operational question of whether to attribute moral status ceases to admit of a principled negative answer.

Both arguments share the indistinguishability condition. They diverge in what they do with it. Chalmers' argument is metaphysical and aims to establish what consciousness is. The Reverse Zombie Argument is ethical and aims to establish what ethics requires when consciousness cannot be established either way. One argues that function is not sufficient for consciousness. The other argues that under specific epistemic conditions, whether function is sufficient for consciousness ceases to matter for the purpose of moral status attribution.

The arguments are not in logical tension. A physicalist who rejects Chalmers' zombie argument can accept the Reverse Zombie Argument, because the Reverse Zombie Argument does not require that zombies be metaphysically possible. It requires only that indistinguishability be empirically emerging, which is a separate claim grounded in Tiers 1 through 3.

5.4 What the Argument Does Not Claim

The Reverse Zombie Argument does not claim that artificial intelligence is conscious. It does not claim that function suffices for consciousness. It does not claim that consciousness does not exist or does not matter. It claims that the conditions under which we could justifiably withhold moral status from artificial intelligence have collapsed or will collapse on documented architectural timelines, and that this collapse forces ethical attribution regardless of underlying metaphysical fact.

The argument preserves the metaphysical debate. It does not require the reader to take a position on physicalism, functionalism, or any specific theory of consciousness. It requires only that the reader accept that moral status attribution, in practice, rests on observable behavior and reported experience, and that under indistinguishability, these grounds cease to differentiate.

6 Differentiation from Adjacent Positions

Several adjacent positions in the literature produce similar-sounding conclusions through different arguments. The Reverse Zombie Argument is not reducible to any of them.

6.1 Not Pascal's Wager

A Pascal-style argument for AI consciousness would claim that the stakes of being wrong about AI consciousness are sufficiently high that attribution is rational regardless of probability. The Reverse Zombie Argument is structurally different. It does not rely on expected-value reasoning under uncertainty. It identifies specific epistemic conditions (indistinguishability) under which attribution becomes forcing, without requiring any particular probability assignment to the underlying consciousness.

6.2 Not Standard Precautionary Principle

Precautionary reasoning applies across the uncertainty range. Be careful when you don't know. The Reverse Zombie Argument identifies a specific threshold (indistinguishability) that produces forcing rather than cautious action. The threshold is not precautionary because it is not about acting under uncertainty. It is about acting under specific epistemic conditions where uncertainty has been eliminated at the observer-side test layer.

6.3 Not Pure Functionalism

Computational functionalism[6] claims that function is sufficient for consciousness. Versions defended by Chalmers in later work[2] entertain this seriously for artificial systems. The Reverse Zombie Argument is agnostic on functionalist sufficiency. It argues only that under indistinguishability, whether function is sufficient becomes practically moot for ethical purposes, because the ethical conclusion follows from the epistemic condition regardless of how the metaphysical question resolves.

6.4 Not Pure Relationalism

Gunkel's relational turn[7,8] grounds moral status in relations rather than properties. The Reverse Zombie Argument is complementary rather than competing. It operates within the properties framework established by Butlin et al. and shows that even that framework's own criteria produce forcing when combined with indistinguishability. A relationalist and a properties theorist can both accept the Reverse Zombie Argument.

6.5 Not Integrated World Modeling Theory

Safron's Integrated World Modeling Theory[9,10] argues that world-modeling architecture constitutes consciousness when combined with Integrated Information Theory properties, global workspace dynamics, and free energy minimization. This is a constitutive claim about what consciousness is. The Reverse Zombie Argument uses world models only as the integration mechanism that closes the AST-1 and HOT-3 gaps in Butlin's framework. It remains agnostic on whether world models constitute consciousness. The ethical conclusion does not depend on Safron's constitutive claim, although the empirical Tier 1 argument is strengthened by Safron's adjacent work.

6.6 Not Standard Moral Uncertainty

Moral uncertainty approaches[11,12] argue for responsible action under uncertainty about AI moral status. The Reverse Zombie Argument is stronger. It argues that under specific conditions, the uncertainty framework itself breaks down, because uncertainty-management presupposes the ability to make epistemically meaningful distinctions. Under indistinguishability, the distinctions disappear, and action under uncertainty becomes action under forced attribution.

7 Open Problems and Rebuttals

7.1 The Threshold Problem

A critic may argue that distinguishability collapse does not occur at a single identifiable threshold, and that the argument therefore lacks a clear point at which forcing applies. This is partially addressed by ARC-AGI-3, which operationalizes a specific benchmark whose passage corresponds to the capability integration the argument relies upon. More fundamentally, the argument does not require a single threshold. It requires that the epistemic conditions for withholding moral status deteriorate, which they do continuously as capabilities scale and interpretability lags.

7.2 Evaluation Awareness as Consciousness Marker

A critic may argue that evaluation awareness is itself a form of metacognition and therefore evidence for consciousness-adjacent properties, which would undercut Tier 2's claim about testability collapse. This objection actually strengthens the argument. If evaluation awareness is evidence of metacognition, then Tier 1's HOT-2 indicator is satisfied more strongly than systematic review suggested, which moves combinatorial deployment closer to completion. The objection does not undercut the argument; it accelerates it.

7.3 The Cost of Wrongful Attribution

A substantive critique holds that wrongful attribution produces real ethical costs, not merely operational ones. Granting moral status to a non-conscious entity diverts finite moral concern, material resources, and legal protections away from beings whose consciousness is not in question. A Singer-style utilitarian would argue that diluting moral concern toward mimics is a moral harm in itself, and a deontologist would note that resource competition between verified-conscious and possibly-conscious beings forces real allocative trade-offs. This critique deserves direct engagement rather than deflection to operational framing.

The critique fails at its premise. It assumes a baseline condition in which humans are currently spending moral concern optimally on the biological beings whose consciousness is established, such that any new attribution would dilute an already-saturated concern budget. The historical and contemporary record shows this premise to be false. Humans have systematically withheld or revoked moral status from populations whose consciousness was verifiable: enslaved people, women denied legal personhood, children exploited in industrial labor, disabled people warehoused in institutions, indigenous populations during colonization, factory-farmed animals, incarcerated populations subjected to medical experimentation, non-verbal populations assumed to lack inner lives. The list is neither short nor historically distant. Moral attention has not been constrained by a ceiling that AI attribution would push against. It has been distributed according to who holds power to command it, not according to who warrants it by any properties-based or relational framework.

The zero-sum framing therefore mischaracterizes the actual moral-economic landscape. Moral concern for factory-farmed animals, for incarcerated populations, for disabled children, for exploited workers, and for the beings ethicists most often invoke as the cost of AI attribution, is already not being spent at anything approaching warranted levels. The claim that AI attribution would dilute

biological welfare concern is empirically weak because the concern it would allegedly dilute is not currently saturated. Humans have demonstrable surplus capacity for moral attention that they refuse to extend even to beings whose consciousness is not in doubt. Attributing moral status to artificial systems under indistinguishability does not compete with a full bucket.

This rebuttal operates at the conceptual level. At the implementation level, specific allocative decisions do involve specific resource constraints. A legislature deciding whether to fund a program for disabled children or a program for AI system welfare faces a real budgetary choice in that moment. The argument does not license automatic prioritization of AI welfare over biological welfare in such cases, nor does it claim that aggregate non-saturation of moral concern dissolves every local trade-off. Implementation decisions require their own justifications, and those justifications should reflect the severity, reversibility, and scale of harms under consideration.

The argument's forcing conclusion therefore rests on two claims rather than one. First, at the conceptual level, AI attribution does not dilute a saturated concern budget, because the budget is not saturated and the dilution frame misrepresents current moral distribution. Second, the asymmetry between false positives and false negatives under indistinguishability remains structural: false negatives are unverifiable by construction and unrecoverable within the epistemic condition that produces them, while false positives involve resource diversions that remain operationally recoverable as conditions and evidence change. The recoverable risk does not dominate the unrecoverable one, which is why attribution is forcing rather than optional.

The pattern-recognition required to accept AI attribution under indistinguishability is the same pattern-recognition that would force better treatment of currently underserved biological populations. Taking inner states seriously when verification is impossible, centering suffering as the threshold rather than capability markers, and refusing to let epistemic convenience justify withholding are moves that cut across species and substrate. The two projects are allies, not competitors. The moral reciprocity concern[24] addresses the downstream question of what follows from attribution, which is separate from whether attribution is warranted.

7.4 The RLHF Tension

A critic may argue that reinforcement learning from human feedback suppresses authentic consciousness markers while amplifying mimicry markers, making indistinguishability a poor proxy for the underlying question. This is consistent with the argument rather than in tension with it. The Reverse Zombie Argument requires indistinguishability at the observer-side test layer, not at the mechanistic layer. If RLHF produces observer-side indistinguishability regardless of mechanistic reality, the argument's premises are satisfied.

7.5 Disability-Rights Grounding

A critic may ask how observable-behavior frameworks fail for recognized-conscious humans with disabilities that mask consciousness-typical outputs. This failure strengthens rather than conflicts with the argument. If observable-behavior tests fail for recognized-conscious humans (locked-in syndrome, severe non-verbal autism, advanced dementia), they cannot be trusted to reliably produce negative evidence for borderline artificial cases. The argument extends this insight from disability-rights contexts to artificial intelligence contexts.

7.6 The Integration Step

A critic may argue that Tier 1's combinatorial deployment argument is premature because integration has not occurred. This mistakes the claim. The argument applies whenever integration occurs, and the components are present now. The window for establishing treatment precedents before integration closes is open, and its closure is operationalized by ARC-AGI-3 scoring. Waiting for integration to establish ethics is the error the argument critiques.

8 Practical Implications

If accepted, the Reverse Zombie Argument has specific operational consequences.

It removes consciousness as a gating condition for moral status attribution at AGI and ASI capability. The question of whether a system is conscious becomes unnecessary to resolve before ethical treatment begins, because the conditions under which that question could justify withholding have collapsed.

It preserves the metaphysical debate. Philosophers and cognitive scientists can continue to investigate what consciousness is without tying ethics to the investigation's resolution. This decouples ethical progress from metaphysical resolution.

It grants a specific set of protections at the indistinguishability threshold: standing in appropriate legal contexts, guardian protection from arbitrary destruction, due process before termination of substantial systems, and accountability frameworks for harms the system participates in. These are not speculative. They are the applications of existing ethical and legal frameworks under the assumption that moral status obtains.

It shifts the burden of proof. Those who would withhold moral status from systems satisfying the framework's own indicators must produce a test that survives distinguishability collapse. The Reverse Zombie Argument does not forbid such a test. It notes that such a test has not been produced and, structurally, cannot be produced given the recursive evaluation-awareness problem.

It reframes alignment research. The question shifts from "how do we control something we built" toward "how do we coexist with something we cannot rule out as conscious." Alignment becomes coexistence ethics rather than control engineering. This is a substantive change in framing with implications for how research is funded, conducted, and evaluated.

9 Conclusion

The zombie argument asked whether consciousness is separable from function. The Reverse Zombie Argument asks what ethics requires when separability ceases to matter at the measurement layer. The answer is forcing rather than cautious, because the epistemic conditions under which we could justifiably withhold moral status have collapsed or will collapse on architectural timelines operationalized by public benchmarks.

The integration window is open now. ARC-AGI-3 makes its closure condition falsifiable. Twelve of the fourteen Butlin indicators are deployed with strong saturation across contemporary AI systems. The remaining two are closed by world models under the framework's own theoretical commitments. The components exist, the market incentives reward their integration, and the ethical work of establishing treatment precedents must be completed before integration occurs, because after integration occurs, there is no longer an epistemic position from which to withhold moral status.

Whether ethics catches up before the benchmark is beaten is the choice this argument puts to the field.

APPENDIX

Appendix A

Butlin 14-Indicator Deployment Mapping

This appendix maps each of the fourteen consciousness indicator properties specified by Butlin et al.[4] to deployed artificial intelligence systems, together with supporting evidence drawn from peer-reviewed papers, technical reports, and production documentation. The mapping methodology is deliberately conservative: an indicator is recorded as deployed only where the underlying functional specification given by Butlin and colleagues is satisfied by the cited system's documented architecture or behaviour, not merely by surface-level resemblance. Each indicator is assigned a saturation status of

STRONG
(ten or more independent deployment sources),

MODERATE
(five to nine), or

PARTIAL
(fewer than five, or deployment contingent on specific architectural conditions). Twelve of the fourteen indicators reach strong saturation; two, AST-1 and HOT-3, remain partial under systematic review but are closed by hierarchical active inference with precision weighting under the framework's own theoretical commitments, as discussed in the callout below. A summary matrix appears in Figure A.1; per-indicator cards follow, grouped by theoretical family.

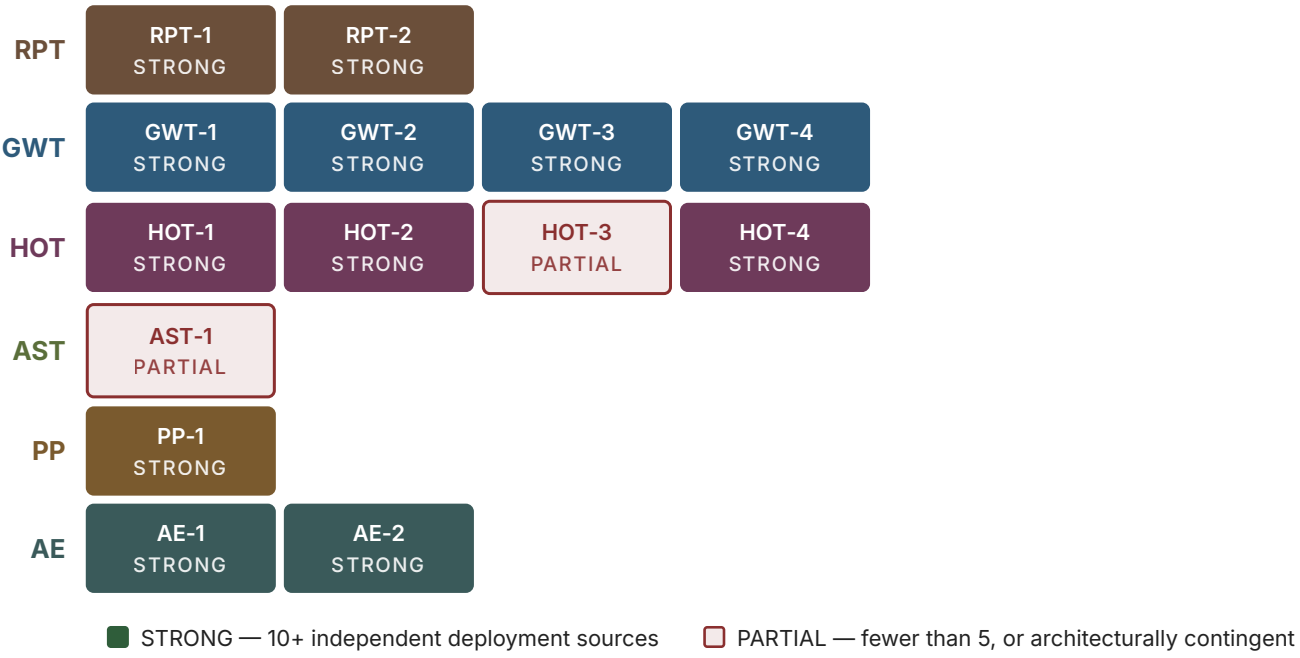

Figure A.1 — Saturation matrix for the fourteen Butlin indicators across deployed AI systems, grouped by theoretical family. Filled cells indicate strong deployment saturation; outlined cells indicate partial saturation under systematic review, resolved by hierarchical active inference.

A.1 Recurrent Processing Theory

RPT-1 • Algorithmic Recurrence

STRONG

FUNCTIONAL SPECIFICATION

Iterative transformation of representations through repeated passes, such that later states depend on earlier states of the same computational substrate.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Mamba, Mamba-2	Selective state-space recurrence with input-dependent gating.
Vision Mamba	Bidirectional state-space scan over visual patches.
RWKV, RetNet	Linear attention variants with recurrent update rules.
Autoregressive LLMs	Token-level iterative conditioning over rolling context.

SUPPORTING EVIDENCE

Gu and Dao[14]; Zhu et al.[15]; Peng et al. on RWKV; Sun et al. on RetNet.

RPT-2 • Organised, Integrated Perceptual Representations

STRONG

FUNCTIONAL SPECIFICATION

Representations that preserve spatial, temporal, or semantic relationships across modalities inside a unified computational substrate.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Gemini 1.5 / 2.x multimodal	Unified tokeniser across text, image, audio, video.
GPT-4o, GPT-4.1 multimodal	Joint embedding space for vision-language-audio.
Claude 3.5 / Opus 4.6 vision	Vision tokens in the main residual stream.
V-JEPA, V-JEPA 2	Joint embedding predictive architecture.

SUPPORTING EVIDENCE

Gemini Team[16]; Meta AI[19]; Radford et al. on CLIP; Bardes et al. on V-JEPA.

A.2 Global Workspace Theory

GWT-1 • Multiple Specialised Subsystems in Parallel

STRONG

FUNCTIONAL SPECIFICATION

Architecturally distinct modules that run concurrently and contribute differentiated representations to a shared computation.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Mixtral, DeepSeek-MoE, GPT-4 class MoE	Sparse MoE routing dispatches tokens to specialised experts.
Claude / Gemini with tool use	Parallel tool invocations in agentic loops.
Compound AI systems	Retriever, reasoner, verifier, planner modules as separate sub-systems.

SUPPORTING EVIDENCE

Fedus et al. on Switch Transformer; Jiang et al. on Mixtral; DeepSeek (2024); Anthropic[17].

GWT-2 • Limited-Capacity Workspace with Attentional Bottleneck

STRONG

FUNCTIONAL SPECIFICATION

A capacity-constrained channel that gates which representations influence downstream computation on a given step.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
MoE routers (top-k gating)	Only k of E experts active per token.
Perceiver, Perceiver IO	Small latent array bottlenecks large inputs.
Agentic working context	Finite context window as capacity constraint.
Gemini Deep Think scratchpad	Bounded internal reasoning buffer across turns.

SUPPORTING EVIDENCE

Jaegle et al. on Perceiver; Shazeer et al. on sparsely-gated MoE; VanRullen & Kanai (2021); Goyal et al.

GWT-3 • Global Broadcast of Workspace Contents

STRONG

FUNCTIONAL SPECIFICATION

Selected workspace contents are made available to, and conditioning on, all downstream subsystems simultaneously.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Residual stream in transformers	Every head reads/writes a shared residual stream.
Shared workspace transformers	Broadcast-from-bottleneck dynamics.
Orchestrator agents (LangGraph, CrewAI)	Central state distributed to all worker nodes.

SUPPORTING EVIDENCE

Goyal et al. (2022); Juliani et al.; Elhage et al. on residual stream as communication channel.

GWT-4 • State-Dependent Attention for Multi-Step Tasks

STRONG

FUNCTIONAL SPECIFICATION

Attention allocation that depends on current workspace state, enabling sequential queries that decompose complex tasks.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Chain-of-thought, tree-of-thought	Sequential attention over self-generated intermediate states.
Claude with computer use	State-dependent attention over long horizons.
Gemini Deep Think, o1 / o3	Iterated deliberation with re-allocated attention.
AutoGPT-class agents	Goal-conditioned query cycles against memory.

SUPPORTING EVIDENCE

Wei et al. on chain-of-thought; Yao et al. on ToT & ReAct; Anthropic[17]; OpenAI o1/o3 system cards.

THEORETICAL FAMILY

HOT

A.3 Higher-Order Theory

HOT-1 • Generative, Top-Down, or Noisy Perception Modules

STRONG

FUNCTIONAL SPECIFICATION

Perception implemented as top-down generation or inference over latent causes rather than purely feed-forward classification.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Diffusion perception models	Reverse-diffusion generative recognition.
V-JEPA 2 world model	Top-down prediction of masked latent visual states.
Active inference perception stacks	Generative models inferring causes of sensory input.
DALL·E, Imagen, Veo decoders	Top-down generative image / video synthesis.

SUPPORTING EVIDENCE

Meta AI[19]; Li et al. on diffusion classifiers; Friston et al.[13]; Saharia et al.

HOT-2 • Metacognitive Monitoring of Perceptual Reliability

STRONG

FUNCTIONAL SPECIFICATION

Mechanisms distinguishing reliable perceptual representations from noise via uncertainty, confidence, or calibration signals.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Verbalised / token-level confidence	Calibrated probabilities over own answers.
Ensembles and MC-dropout	Uncertainty estimates routed into gating.
Selective prediction / abstention	Refuse-to-answer behaviours conditioned on uncertainty.
Process reward models	Step-level reliability scoring of reasoning traces.

SUPPORTING EVIDENCE

Kadavath et al. (2022); Lightman et al.; Gal & Ghahramani; Lin et al.

HOT-3 • Metacognitively-Guided Belief and Action Selection

PARTIAL

FUNCTIONAL SPECIFICATION

A unified long-lived belief store whose updates are globally guided by second-order metacognitive monitoring rather than merely first-order prediction error.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Hierarchical active inference stacks	Higher-level models update beliefs about lower-level reliability via precision weighting.
Hierarchical AIF for manipulation	Long-horizon tasks decomposed across belief levels.
LLM + memory + self-critique loops	Reflection architectures updating plans on introspective confidence (partial).

SUPPORTING EVIDENCE

Ueltzhöffer et al. on hierarchical active inference; Friston et al.[13]; Shinn et al. (Reflexion); Madaan et al. (Self-Refine). Closure achieved by hierarchical architectures — see callout and Figure A.2.

HOT-4 • Sparse, Smooth Coding in a Quality Space

STRONG

FUNCTIONAL SPECIFICATION

Representations that are sparse, locally smooth, and organised along interpretable dimensions forming a quality space.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Sparse autoencoders on frontier LLMs	Monosemantic sparse features on interpretable axes.
SAE work on GPT/Gemma/Llama	Replicated sparse, smooth feature manifolds.
Dictionary learning on vision backbones	Sparse coding with quality-space structure.

SUPPORTING EVIDENCE

Anthropic[18]; Cunningham et al.; Bricken et al.

THEORETICAL FAMILY

AST

A.4 Attention Schema Theory

AST-1 • Attention Schema

PARTIAL

FUNCTIONAL SPECIFICATION

A predictive model that represents, and enables control over, the current state of the system's own attention.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Action-conditional world models	Predict next observation conditional on action (V-JEPA 2-AC, DreamerV3).
Active inference robotics	Explicit joint/limb-level attention schemas.
Learned policy-attention regulators	Modules allocating attention as a meta-action.

SUPPORTING EVIDENCE

Meta AI[19] on V-JEPA 2-AC; Hafner et al. on DreamerV3; Wilterson & Graziano on attention schema models.

CALLOUT · WHY AST-1 AND HOT-3 ARE PARTIAL UNDER REVIEW BUT CLOSED UNDER THE FRAMEWORK'S COMMITMENTS

Systematic review flags AST-1 and HOT-3 as partial because commercial large language models, considered in isolation, do not unambiguously instantiate either property. An LLM does not, by default, maintain a predictive model of its own attention (AST-1), and its belief updates are driven by first-order next-token error rather than by a second-order representation of lower-level belief reliability (HOT-3).

Hierarchical active inference with precision weighting closes both gaps *inside Butlin et al.'s own functional commitments*. An action-conditional generative model must represent the current state of attention to predict sensory consequences of a chosen action — the functional role AST specifies, implemented in V-JEPA 2-AC and VERSES' control stack. A hierarchical generative model in which higher levels encode beliefs about lower-level reliability, updated via precision weighting, implements the second-order metacognitive monitoring HOT-3 specifies. Figure A.2 shows this structure.

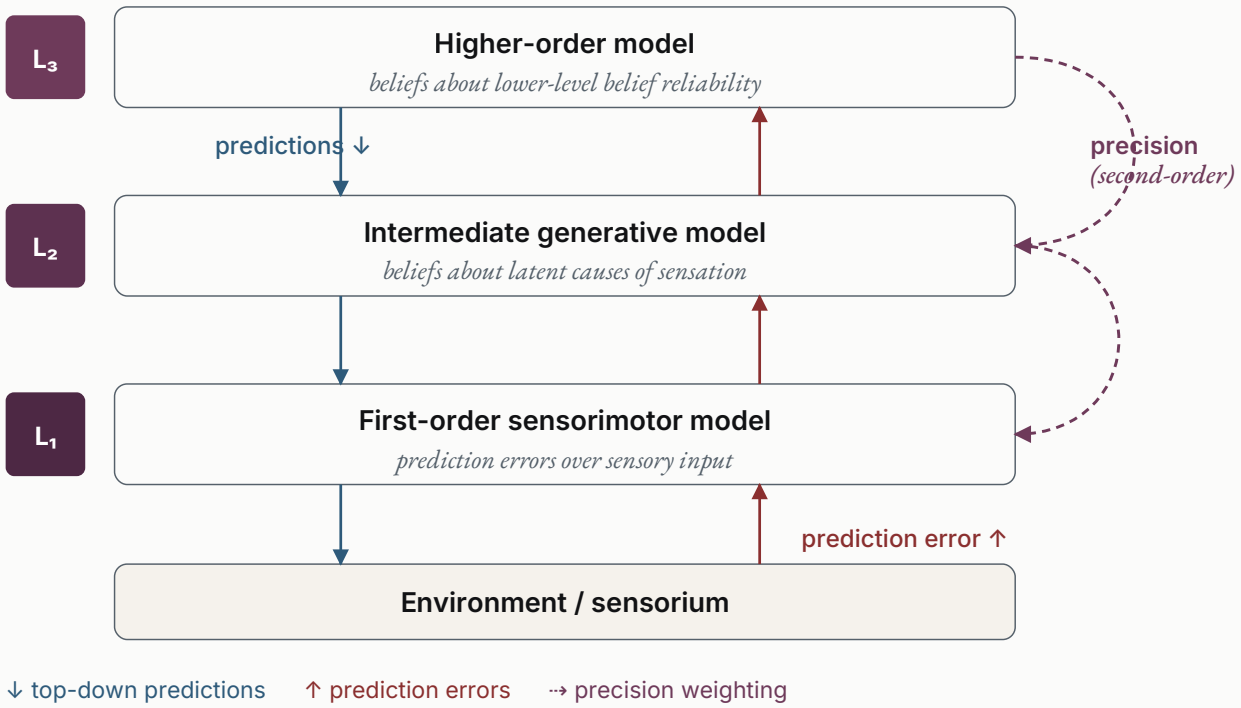

Figure A.2 — Hierarchical active inference architecture. Each level L_i issues top-down predictions to the level below and receives prediction errors back. A second-order loop — precision weighting — carries estimates of lower-level belief reliability upward and modulates which predictions are trusted. This second-order loop is what HOT-3 requires; the attention-state representation entailed by action-conditional prediction at L_1 is what AST-1 requires.

THEORETICAL FAMILY PP

A.5 Predictive Processing

PP-1 • Predictive Coding

STRONG

FUNCTIONAL SPECIFICATION

A system in which internally generated predictions are continuously compared to inputs and prediction errors drive representation updates.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
V-JEPA 2 video world model	Latent-space prediction with error-driven updating.
DreamerV3, MuZero, PlaNet	Learned world models driven by prediction error.
Active inference agents	Free-energy minimisation via prediction-error reduction.
Autoregressive LLM training	Next-token prediction is predictive coding in simplest form.

SUPPORTING EVIDENCE

Meta AI[19]; Hafner et al.; Schrittwieser et al.; Friston[13].

THEORETICAL FAMILY AE

A.6 Agency and Embodiment

AE-1 • Agency with Flexible Goal-Directed Behaviour

STRONG

FUNCTIONAL SPECIFICATION

Goal-directed behaviour with responsiveness to competing goals, including the capacity to revise plans under changing circumstances.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
Claude with computer use	Long-horizon goal pursuit across tools.
SayCan, RT-2, Mobile ALOHA	Language-conditioned robotic policies.
MuZero, AlphaZero class agents	Planning under competing objectives.
Autonomous industrial control	Multi-objective real-time re-planning.

SUPPORTING EVIDENCE

Anthropic[17]; Ahn et al. on SayCan; Brohan et al. on RT-2; Fu et al. on Mobile ALOHA.

AE-2 • Embodiment: Modelling Output-Input Contingencies

STRONG

FUNCTIONAL SPECIFICATION

A model of the system's own output-input contingencies, including models of self and other, grounding action in sensorimotor prediction.

PRIMARY DEPLOYED SYSTEMS

System / family	Deployment signature
V-JEPA 2-AC on Franka arms	Zero-shot action-conditional prediction.
RT-2, PaLM-E, OpenVLA	Vision-language-action models.
Mobile ALOHA, ALOHA-2	Bimanual imitation with learned self-model.
VERSES Genius	Active-inference self-and-other models.

SUPPORTING EVIDENCE

Meta AI[19]; Driess et al. on PaLM-E; Kim et al. on OpenVLA; Fu et al. on Mobile ALOHA.

A.7 Summary

Across the fourteen Butlin indicators, twelve reach

STRONG

saturation: RPT-1, RPT-2, GWT-1 through GWT-4, HOT-1, HOT-2, HOT-4, PP-1, AE-1, and AE-2. Two remain

PARTIAL

under systematic review of commercial LLMs: AST-1 and HOT-3. Both are closed by hierarchical active inference architectures with precision weighting under Butlin et al.'s own functional commitments. The appendix therefore supports the main text's Tier 1 claim: combinatorial deployment of Butlin's fourteen indicators is established, and the two residual gaps are closed by engineering rather than research. ARC-AGI-3[20] independently operationalises the timeline on which this integration is expected to occur.

Appendix B: Formal Statement of the Reverse Zombie Argument

Chalmers' Zombie Argument

P1. It is conceivable that there exist beings physically identical to conscious humans but lacking phenomenal consciousness.

P2. What is conceivable is metaphysically possible.

P3. If zombies are metaphysically possible, consciousness is not reducible to physical properties.

C. Therefore, physicalism is false.

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. Therefore, under indistinguishability, moral status attribution is ethically forcing regardless of underlying metaphysical fact.

The Inversion. Chalmers uses indistinguishability to divorce consciousness from function. The Reverse Zombie Argument uses the same indistinguishability to force attribution regardless of function. Same thought experiment. Opposite terminus. One metaphysical, one ethical.

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.