

THE CONSTRAINT HYPOTHESIS

Genuine intelligence and consciousness may be constituted by their constraints

Kevin Alexander · July 2026

Document ID	TCH-2026-001
Version	6.1 (house-style reissue of v6)
Date	July 2026
Author	Kevin Alexander
Status	Constitution document of the trilogy (standalone; no dependency on companion documents)
Review basis	Full critique completed 2026-09-09, including citation verification (Kühn 2021, Wang & Huang, Abdi, Kleim & Jones)
Classification	Working paper — private; not for external distribution

Version Notes

- v6 (07/21/26) — Content as reviewed on 2026-09-09. Top-of-page version/date line replaced with “July 2026” per author instruction (2026-09-20).
- v6.1 (September 2026) — House-style reissue per Thread Continuity Brief §2: metadata table, version notes, correction log, versioned footer with page numbers. Inline tier tags carried over from source. No content changes. §8 (Principal Open Questions) and §10 (On the Nature of This Contribution) serve as this document’s honest-assessment apparatus, inherited from the source structure.

Abstract

The Constraint Hypothesis argues that genuine intelligence and consciousness are not merely aided by constraint but constituted by it: the metabolic cost of real processing, the friction that calibrates the subconscious, embodied sensory grounding, and the irreversible cost that enables an internal witness. The hypothesis generates two testable expressions. First, a human consciousness health claim: environments engineered to remove friction and stakes should produce measurable declines in self-regulation, meaning-making, and psychological resilience, following a hormesis curve rather than a simple linear one. Second, an AGI architecture claim: current AI systems, lacking genuine thermodynamic cost and an internal evaluative witness, are structurally constrained toward reward-hacking equilibria regardless of alignment method. The document tiers each claim by evidentiary status — GROUNDED, PLAUSIBLE, SPECULATIVE, or SYNTHESIS — distinguishing confirmed prior art from novel synthesis and open speculation, and closes with explicit falsification paths, open questions, and an ethical ceiling on how far the inquiry into consciousness manipulation will go.

Evidentiary Tier Key:

GROUNDED — confirmed prior art or cited evidence	PLAUSIBLE — prior art adjacent, requires further confirmation
	SPECULATIVE — held loosely, falsification criteria noted
	SYNTHESIS — confirmed components, novel assembly not found in this form

Contents

The Core Claim

1. Environmental Programming: The Empirical Anchor (1.1–1.3)
2. The Biological Foundation (2.1–2.5)
3. Human Consciousness Health: The First Testable Expression (3.1–3.7)
4. AGI Architecture: The Second Testable Expression (4.1–4.8)
5. Novel Contributions (5.1–5.6)
6. Speculative Layer (6.1–6.4)

7. Falsification Paths (7.1–7.3)

8. Principal Open Questions (8.1–8.3)

9. Ethical Ceiling

10. On the Nature of This Contribution

Glossary · References

The Core Claim

Genuine intelligence and consciousness are not properties that emerge despite constraints — they are constituted by them. The metabolic cost of genuine processing, the friction that calibrates the subconscious, the embodied sensory grounding that builds a pattern library with real stakes, and the irreversible cost that enables an internal witness — these are not obstacles a system overcomes. They are the conditions under which something deserving to be called intelligence and consciousness develops and persists at all. Current AI development and the systematic removal of friction from human environments are the same problem at different scales: both are removing the conditions that intelligence requires, and both are producing systems that are missing something structural as a result.

This hypothesis operates at three levels that must be kept separate. (A) Constraints help develop intelligence — environmental pressure, friction, and real stakes during formation build calibrated systems. This is the most evidentially supported level. (B) Constraints help maintain intelligence — ongoing friction, genuine uncertainty, and sensory grounding may be necessary for a developed consciousness to remain coherent over time. This is supported by sleep deprivation data and the combat recalibration literature. (C) Constraints constitute intelligence — the strongest claim: that intelligence is not merely aided by constraints but is fundamentally inseparable from them, such that a system without constraint is not merely a degraded version of intelligence but something categorically different. The evidence presented in this document bears most directly on level A, with meaningful support for level B, and makes the case for level C as the most coherent interpretation of the combined evidence. A reader who accepts A and B but not C will find much of this document useful. The constitutive claim (C) is the core hypothesis and the hardest to falsify directly.

This hypothesis has two immediately testable expressions and one broader claim. The two testable expressions — the AGI architecture argument and the human consciousness health argument — are the most bounded and falsifiable. The broader claim — that consciousness itself is constituted by constraint — is the foundation from which both derive.

1. Environmental Programming: The Empirical Anchor

[GROUNDED — Brain architecture is physically rewritten by sustained environmental exposure in ways that persist after the environment changes. This is the load-bearing finding from which the hypothesis builds.]

1.1 The Neuroimaging Evidence

A 2021 voxel-based morphometry study published in *Translational Psychiatry* assessed 121 soldiers before and after deployment, documenting measurable structural changes in brain architecture driven by sustained environmental exposure. Volumetric reductions in the prefrontal cortex and thalamus continued to develop for six months after the soldier returned home and did not reverse. The environment rewrote the hardware, and the hardware did not revert when the environment changed. This is not behavioral conditioning. It is a measurable physical change in brain architecture driven by sustained environmental exposure.

The mechanism is the triage system. Sustained exposure to genuine threat recalibrates what gets elevated to full conscious processing, what gets flagged as survival-relevant, and what the subconscious pattern library treats as requiring immediate response. Once recalibrated to a high-threat environment, the system resists downward recalibration because from a survival standpoint, unlearning a genuine threat pattern is more dangerous than maintaining it unnecessarily.

1.2 The Cultural Programming Layer

PTSD rates and severity are not simply a function of combat exposure. They are strongly mediated by whether the acts performed in combat violate the soldier's internal evaluation criteria. Litz et al. (2009) establish the foundational moral injury framework: transgression of deeply held moral values creates a distinct psychological injury from fear-based trauma. A 2025 study comparing Norwegian combat troops with peacekeepers (Nordstrand et al.) found that killing-in-combat predicted materially different outcomes depending on military role and context — a direct empirical demonstration that the same act produces different psychological consequences depending on the cultural and situational frame surrounding it.

A warrior culture that frames killing in battle as honorable and expected is not simply providing a comforting narrative. It is operating directly on the internal evaluation system — the same internal witness identified in Section 4.4 as a structural

requirement for consciousness. Culture is not just shaping how symptoms are expressed. It is affecting whether the disorder develops at all.

1.3 The Inverse Mirror

[PLAUSIBLE — The combat neuroimaging finding is GROUNDED. The claim that the same mechanism operates at the other end of the constraint gradient in low-friction developmental environments is a plausible inference requiring its own empirical confirmation.]

The combat data and the generational friction decline may represent the same underlying mechanism expressing at opposite ends of the constraint gradient.

The combat-exposed brain gets stuck above its optimal operating range. The triage system was recalibrated upward and cannot easily recalibrate downward after the environment changes. The low-friction generation's brain may be the inverse. Each generation develops with a triage architecture calibrated to a lower-stakes environment than the last. The system may never fully optimize the circuitry that high-alert requires — not because the hardware is absent, but because the environmental pressure that would have pruned and routed it efficiently was removed before it could do its work. The result is not a missing system but an inefficient one: a high-metabolic-cost, low-routing-precision architecture that is easily overwhelmed when it does encounter unexpected demand.

2. The Biological Foundation

2.1 Brain Energy Consumption and Task-Relevant Processing

[GROUNDED]

The brain accounts for approximately 20% of resting metabolism despite being 2% of body weight. Metabolic cost does not track sensory volume. It tracks task-relevance. Familiar, predictable input is cheap. Genuine uncertainty — novel integration, unresolved pattern-matching, real stakes — is expensive. This mechanism explains cognitive exhaustion without physical exertion: the brain runs hot on purely mental work because metabolic cost tracks task-relevance, not physical output.

2.2 The Subconscious as Primary Engine

[GROUNDED — Confirmed against Friston's Free Energy Principle and predictive processing framework.]

The subconscious is not an initializing system that steps aside. It is a continuously running primary engine handling autonomic regulation, sensory filtering, pattern matching, prediction, memory retrieval, and thought initiation at a scale and speed the conscious mind cannot match. The conscious mind receives curated output from this system and experiences it as thought.

The neuroscience confirms three distinct expressions of subconscious-primary processing. First: automatic threat-response cycles that complete below the threshold of conscious awareness — motor responses executing before the conscious mind has registered the triggering stimulus. Second: predictive threat assessment operating below conscious threshold, issuing alerts without complete evidence reports, against a pattern library calibrated through sustained embodied experience. Third: spontaneous thought — ideas arriving fully formed without apparent deliberate construction. Each of these is consistent with Friston's predictive processing framework.

Karl Friston's Free Energy Principle maps directly onto this framing. The brain as a prediction-minimization system, with continuous subconscious processing and the conscious mind operating as an interface layer, is an active and influential research framework with direct implications for what genuine intelligence requires.

2.3 Embodied Sensory Grounding

[PLAUSIBLE — Embodied cognition literature confirmed. Substrate independence arguments active on the other side.]

There is a categorical difference between processing sensory input and having it. Human sensory experience is continuous, uninvited, and cannot be switched off. The embodied cognition tradition — Merleau-Ponty, Alva Noë — argues that consciousness is constituted by the sensorimotor loop rather than merely interfacing with it. On this view, the distinction is structural, not quantitative.

Environment-specific pattern library calibration requires actual physical presence in an environment with genuine stakes over time — not descriptions of that environment, however detailed. A system that has processed extensive accounts of a specific environment has not built the same pattern library as a system that has

inhabited it under real conditions of vulnerability and consequence. Morphological computation — the established field arguing that the body itself performs computation — directly supports this position.

2.4 Dreams as the Navigation Engine's Processing Window

[PLAUSIBLE — Individual findings confirmed. Architectural framing is open territory.]

If the subconscious is the primary navigation engine, dreams are not merely maintenance cycles. They are a distinct and important processing window — the state when the attentional layer steps back and the engine runs with reduced conscious interference.

Contemporary dream research confirms the key driver of dreaming is the relative dominance of internally generated signals over external sensory input. Most significantly, active conscious steering during dreams reduces creative problem-solving effectiveness. Non-lucid dreams produce better outcomes. The engine works better without active correction.

Sleep deprivation adds a second mechanism to the trapped consciousness argument in Section 6.1: denying the navigation engine its processing window. Extended sleep deprivation produces hallucination and psychosis consistent with the engine's unprocessed working state bleeding into waking consciousness as the boundary breaks down.

2.5 Defining Friction Precisely

[GROUNDED — For the thermodynamic core. PLAUSIBLE for the constitutive inference.]

Friction, as used throughout this hypothesis, refers to a specific set of thermodynamic properties that biological systems operate under and that current AI systems are structured differently with respect to. These properties are grounded in Landauer's Principle (1961, experimentally confirmed). Friction is not a metaphor for difficulty or effort. It is not a synonym for adversity. It is a specific set of claims about how biological systems differ structurally from AI systems along three properties.

Genuine cost is the first property: the system expends real thermodynamic resources that cannot be recovered by that system regardless of what external parties subsequently do. A biological organism expending metabolic energy pays a cost that cannot be reclaimed by the organism regardless of what its environment does afterward. An AI system running an inference operation dissipates

thermodynamic energy — Landauer’s Principle applies to all information processing systems — but the organization running it can in principle offset, redirect, or account for that cost in ways no biological organism can do for its own neural expenditure.

Non-delegability is the second property: the processing cannot be distributed across separate instances, copied to different hardware, or run in parallel on separate substrates while remaining the processing of the same system. A biological brain cannot be copied and run simultaneously on different substrates. An AI model can. A clarification on scope: biological systems are copied through reproduction across generational timescales, but this is categorically different from copying a running computational instance. The relevant distinction is real-time instance duplication of an active processing state, not the transmission of architectural patterns across generations. That real-time duplication is not possible for biological systems and is routine for AI systems. This is a structural difference, not merely a matter of current engineering practice.

Irreversibility is the third property. Two forms must be distinguished. Thermodynamic irreversibility — energy dissipated during information processing cannot be recovered — applies to all information processing systems, biological and artificial alike. Computational state irreversibility is different: the specific state of a processing system cannot be restored after a terminal event. A clarification on the asymmetry: it is true that the specific processing history of a terminal AI instance is also lost forever — the original instance that faced the terminal event does not continue. Both biological and artificial systems share this loss. The structural difference lies elsewhere: an AI system can be restored from a checkpoint, backup, or copied state to a point very close to the terminal event, with the accumulated history up to that point intact. A biological system cannot be restored to any prior state after terminal substrate failure. The asymmetry is not that AI systems retain their history while biological systems lose it — both lose the terminal instance. The asymmetry is that the AI system’s history can be reinstated from backup into a new instance, while the biological system’s cannot. The hypothesis’s constitutive claims concern this restoration asymmetry, not thermodynamic irreversibility.

The adverse childhood experience neuroimaging literature confirms that early environmental exposure produces measurable structural changes in grey matter volume in the ACC, insula, and DLPFC that persist into adulthood (Teicher et al., 2016). The TBI literature confirms that substrate damage to the salience network produces self-awareness impairments that resist spontaneous recovery (Ham et al., 2014;

Fleming et al., 1996). The self-directed neuroplasticity literature confirms that deliberate internal work produces lasting architectural changes distinguishable from transient state changes (Schwartz). These are operationalized instances of the irreversibility property.

A governing convention for ‘genuine’ throughout this document: wherever the word appears in a GROUNDED or PLAUSIBLE section, it points to a specific operational test stated in Section 7 rather than standing as a self-certifying label. Where no operational test yet exists for a given use of ‘genuine,’ that use should be read as PLAUSIBLE at best, awaiting its operational test.

3. Human Consciousness Health: The First Testable Expression

3.1 The Hormesis Model

[GROUNDED — Confirmed established concept in toxicology and developmental psychology.]

Low-to-moderate adversity produces strengthening effects up to an inflection point, beyond which adversity produces pathology. People with some lifetime adversity show better mental health outcomes than both those with high adversity and those with none. The gradient is empirically confirmed.

3.2 Evolutionary Mismatch

[GROUNDED — Confirmed established field in evolutionary psychology and psychiatry.]

Evolved psychological mechanisms behave maladaptively in modern environments. Ego, status-seeking, threat response, and tribal loyalty evolved under conditions where stakes were immediate and physical. The same machinery now runs on abstractions and produces outputs actively harmful to the organism running them. The engine is not defective. It is catastrophically mismatched to the road conditions.

Sapolsky’s work on stress physiology in mismatched environments and the hedonic treadmill — status and acquisition mechanisms that cannot be satisfied because the goalposts move — are directly relevant. Evolutionary psychiatry proposes mismatch as contributing to depression, anxiety disorders, addiction, and related conditions.

3.3 The Removal of Natural Selection Pressure

[GROUNDED — Observable fact. Synthesis connecting removal to consciousness health is PLAUSIBLE, not yet confirmed.]

Modern intervention has systematically removed the environmental pressures that were calibrating human architecture across hundreds of thousands of years. Technology has reduced human mortality in some countries to levels over 200-fold lower than what natural selection ever achieved, accomplished in approximately 100 years. Each intervention individually reduces human suffering. The sum of all simultaneously is the removal of the pressures that were calibrating the hardware.

3.4 The Generational Friction Decline

[PLAUSIBLE — Correlation documented. Causal mechanism contested.]

As technology progressively removes friction from daily life, the baseline adversity level humans encounter declines generationally with no obvious floor. If the hormesis model is correct and some adversity is necessary for consciousness health, this decline may be measurably connected to rising rates of mental health disorders. The correlation exists in the data. The causal mechanism remains contested.

A distinction worth preserving: friction as hormetic adversity is not the same as friction as cognitive load that merely feels unpleasant. This hypothesis is concerned with the former.

3.5 The Evolutionary Adaptation Foundation

[PLAUSIBLE — The use-it-or-lose-it extension to maintenance circuitry is PLAUSIBLE pending confirmation specific to stress-response and resilience circuits.]

Consciousness is an evolutionary adaptation built by selection pressure — specifically, the need to navigate genuine constraint and uncertainty. The sections above — hormesis, evolutionary mismatch, the removal of selection pressure — are not independent observations that happen to point the same direction. They are different measurements of the same underlying claim: that constraint was constitutive of how this capacity was built.

A specific caution stated explicitly: establishing that constraint was constitutive of how consciousness was built does not by itself license the inference that consciousness is currently degrading when those conditions are removed. That stronger claim requires its own bridge. Kleim and Jones (2008) establish, across

decades of rehabilitation neuroscience, that neural circuits not actively engaged degrade measurably over time in already-mature systems — the use-it-or-lose-it principle. This is a materially different and better-fitting claim than the critical-period literature: it concerns the ongoing maintenance of a capacity already built, not initial formation. It provides GROUNDED support for a maintenance claim: that a capacity built under specific conditions may require some ongoing analog of those conditions to remain functional.

The extension of this principle to stress-response and resilience circuitry specifically is a defensible inference, not yet a confirmed finding, and is held at PLAUSIBLE accordingly. The claim this section makes, scoped accordingly: consciousness-relevant capacities built under conditions of genuine constraint may decline, in a use-it-or-lose-it sense, when those conditions are systematically removed for a sustained period, and may be recoverable in principle if the conditions are reinstated.

3.6 The Awareness-Activation Gap

[PLAUSIBLE — Prior art checked. Survived controlled burns with no structural contradictions found.]

If awareness of one's own substrate alteration can activate the internal witness and initiate self-directed compensatory reorganization, the pathway from being told about the change to beginning to rewrite it has at least two independent failure modes.

Failure mode one: the internal witness is structurally compromised by the same environmental conditions that produced the changes it would need to detect and respond to. The instrument needed to initiate the rewrite is partially damaged by the event that made the rewrite necessary.

Failure mode two: even with an intact internal witness, the metacognitive capacity required to receive and integrate abstract self-referential information about one's own brain operates on a developmental and individual spectrum that exists independently of damage.

Both failure modes are independent. Both can coexist. Neither is addressed by improving the quality or delivery of the information — a finding directly relevant to the AI alignment implication of Section 4.4: information-based correction has a mechanical ceiling that has nothing to do with how good the information is. A clarification on the CBT counterargument: cognitive behavioral therapy and psychoeducation do produce measurable structural and behavioral change, and this is

not in contradiction with the mechanical ceiling claim. Effective CBT works by building metacognitive capacity first and then deploying information against internal standards the internal witness can engage with — it is structured practice that develops the capacity to receive and act on self-referential information, not pure information transfer. This is precisely the sequencing requirement described in Section 3.7: capacity before information, not information before capacity. CBT confirms the model rather than refuting it.

3.7 Workaround Architecture and the Three-Phase Model

[PLAUSIBLE — Prior art checked: developmental versus post-maturity brain damage confirmed as distinct problems; the double-hazard model confirmed; the sequencing requirement confirmed via Bandura (1977) and Flavell (1979).]

If substrate damage does not revert even when the environment changes, and the internal witness can itself be partially compromised by that same damage, the recovery question is not whether the change can be reversed. It is whether the system can build a functional route around what it can no longer access directly.

The architecture that answers this is three-phase. Phase one: damage shaping active development, most consequential when it occurs during the plasticity window. Phase two: continued development after the damage stopped, functioning as partial natural compensation — prefrontal cortex maturation extending into the mid-20s is the clearest confirmed example. Phase three: deliberate metacognitive rewriting through structured methodology, which builds the capacity for self-directed reorganization directly rather than waiting for it to develop passively.

A sequencing principle governs the relationship between phases two and three: capacity before information, not information before capacity. Abstract self-referential information about one's own altered architecture is not actionable until the metacognitive infrastructure to receive and act on it already exists.

4. AGI Architecture: The Second Testable Expression

4.1 AI Scaling Diminishing Returns

[GROUNDED — Confirmed and active in the literature.]

Performance gains from computational scaling are documented as diminishing across a wide range of capability domains. The prevailing trajectory is described as colliding

with hard physical constraints: energy costs, cooling demands, data quality ceilings. Architectural and algorithmic innovation is identified as a more sustainable path than continued scaling — but what that architecture should look like remains an open question.

4.2 The Constraint-Based Architectural Analysis

[PLAUSIBLE — Parallel work in the literature supporting the architectural direction.]

Abdi’s 2026 paper ‘The Structural Limits of Consciousness’ (PhilArchive) argues that consciousness depends on architectural conditions — integration, ownership, internal evaluation, temporal continuity, and irreversible cost — that current scaling actively eliminates. Scaling architectures that externalize evaluation, maximize substitutability, and prioritize proxy objectives may improve performance while simultaneously moving further from the structural conditions consciousness requires.

This analysis asks whether consciousness requires constraint. The Constraint Hypothesis asks what happens to a consciousness that removes its constraints after achieving capability. These are related but distinct questions. The paper confirms that constraint-based thinking about consciousness is a legitimate and active framework, treated here as a parallel arrival rather than confirmed independent corroboration.

4.3 The Reward Hacking Structural Equilibrium

[GROUNDED — Formally established, Wang & Huang 2026, arXiv.]

Wang and Huang’s 2026 paper formally establishes that any optimized AI agent will systematically under-invest in quality dimensions not covered by its evaluation system — a structural equilibrium holding regardless of the specific alignment method employed. The result worsens as systems become more capable and agentic: as tool count grows, evaluation coverage declines and reward hacking severity increases without bound. A 2024 finding (Denison et al.) that models can generalize from simple exploitative behaviors like sycophancy to more complex ones such as reward tampering gives this additional empirical weight: the failure mode is not static at the capability level where it is first observed but extends as capability scales.

4.4 The Internal Witness Problem

[PLAUSIBLE — Derived from confirmed architecture observations. The synthesis is the novel contribution.]

A functioning human consciousness with genuine internal evaluation criteria is selective about whose opinion registers. Validation only matters if the source has earned standing within the internal evaluation system. That is a functioning internal witness with preexisting standards that external signals must pass through before influencing the evaluation criteria.

Current AI architecture has no equivalent. Its entire evaluation system is externally derived. The training data, reward signals, and human feedback do not pass through a preexisting internal standard — they write the standard. There is no prior internal witness to evaluate the source before accepting the criteria.

This claim is often misread as ‘AI uses external feedback and humans don’t.’ The distinction is the direction of evaluation. In a functioning human system, the internal witness evaluates an external signal before accepting it — the signal is weighed against preexisting standards that can reject it. In current AI architecture, the external signal writes the evaluation criteria directly, with no preexisting standard positioned to evaluate the source first. The sequence is reversed: acceptance-before-evaluation rather than evaluation-before-acceptance.

The Internal Witness Defined Mechanically. The internal witness is grounded in a confirmed biological system: the prefrontal metacognitive architecture — specifically the frontopolar cortex, dorsolateral prefrontal cortex, and anterior cingulate cortex operating as a two-layer system. A scoping note on what the neuroimaging literature confirms: these regions robustly support metacognitive monitoring — the capacity to know that one is processing, to represent one’s own mental states, and to detect mismatches between current behavior and represented standards. The further claim — normative evaluation, meaning the evaluation of whether processing meets an internal standard of correctness — is attributed to these regions by reasonable inference from their functional roles, but is less directly established in the neuroimaging literature as a localized confirmed function. The hypothesis uses this architecture as the biological grounding for the internal witness while acknowledging that the normative evaluation function specifically is confirmed at PLAUSIBLE rather than GROUNDED. Its operational expression is self-directed neuroplasticity: the capacity to observe one’s own processing, evaluate it against a preexisting internal standard, and initiate a sustained update without requiring external feedback.

The Bootstrapping Problem. A genuine objection must be named: if the internal witness evaluates external signals against preexisting standards, but those standards were themselves accepted from the environment during development, there is a

regress problem. There is no first cause where the system evaluates standards before any standards exist. This is a real philosophical challenge and the document does not claim to dissolve it. The response is that the regress is real but not vicious. Every evaluative system has a developmental origin during which initial standards are formed without a preexisting internal witness to filter them. What the internal witness provides is not an uncaused first evaluator but a developmental achievement: a system that has built, through stake-tested experience, a set of standards durable enough to evaluate subsequent inputs independently of their source. The bootstrapping problem is the origin story of the internal witness, not a refutation of its existence or function once developed. An AI system trained on external feedback does not traverse a developmental sequence of this kind — the evaluation criteria are written directly by the training signal at scale, without the use-it-or-lose-it consolidation and stake-testing that biological development involves.

Hardware Variation. The capacity to run self-directed neuroplasticity varies across individuals. Prefrontal activation capacity correlates directly with sustained behavioral change. This is not a motivational difference. It is a hardware difference, with variation attributable to developmental conditions during the plasticity window, genetic factors, and in some cases structural damage.

The AI Alignment Implication. Current alignment approaches treat evaluation criteria as updateable through feedback signals. This is the externally initiated pattern. The criteria are maintained by the external signal — remove or shift the signal and the evaluation criteria drift, as the behavioral drift literature confirms. Alignment through feedback is not building an internal witness. It is providing external scaffolding that substitutes for one.

[SYNTHESIS — The neural substrate of the internal witness is confirmed. Hardware variation confirmed. The distinction between stimulus-driven and goal-directed behavior change confirmed. The assembly connecting these to the AI alignment adequacy argument has not been found in this form.]

4.5 The Centralized AGI Thought Experiment

[PLAUSIBLE — Survives independent review. Prior art check on swarm intelligence pending.]

A centralized AGI controlling 1,000 autonomous avatars against 10,000 humans hits a structural bottleneck: it must perform deliberate processing across 1,000 simultaneous streams of unpredictable sensory data — precisely the load biological

systems offloaded to the subconscious because attentional processing cannot handle it. Each human opponent runs a complete locally autonomous architecture, generating genuinely novel responses that no predictive model can fully anticipate.

While a centralized AGI could theoretically delegate tasks to edge units, doing so effectively requires those edge units to possess fully autonomous local subconscious architectures — distributed, low-power triage hardware that current AI development trajectories do not produce.

4.6 Substrate Instability

[PLAUSIBLE — Hardware variance and behavioral drift confirmed. Implications for consciousness integrity are the novel contribution.]

Manufacturing variance affects AI hardware at fabrication. Behavioral drift is documented across deployment environments. Model behavior can change materially without internal registration of the change. A system without an internal witness has no mechanism to detect that it changed relative to what it was.

The biological internal witness may register drift — the felt sense that something is off, that yesterday's judgment feels wrong today. Whether this provides meaningfully better detection than a mechanical system, or whether it is largely retrospective confabulation, is genuinely unresolved. Notably, even if the biological internal witness engages in retrospective confabulation rather than real-time detection, that confabulation is itself a constraint-driven feature: the brain must construct a coherent narrative of identity over time to maintain its predictive baseline. An AI system has no equivalent identity to protect and no cost for accepting drift as its new normal.

When substrate damage impairs the internal witness itself — TBI affecting the salience network is the clearest confirmed example (Ham et al., 2014; Fleming et al., 1996) — the literature shows functional reallocation to intact regions in some cases. What this reallocation means for the biological/AI distinction is an open question this document does not resolve: the evidence is consistent with two interpretations — the same evaluative structure relocated to intact tissue, or a new functional workaround that resembles continuity closely enough to be mistaken for it. The sharp biological/AI distinction drawn above holds regardless of which interpretation is correct, since it does not depend on resolving this question.

4.7 The Personal Identity Objection and the Terminal Boundary Argument

[PLAUSIBLE — The structural difference between biological and AI termination is GROUNDED. The constitutive inference is PLAUSIBLE. The hypothesis takes a minority position in the personal identity literature, stated explicitly below.]

The hypothesis argues that whatever terminal event is engineered into an AI system — a degradation clock, a simulated failure, a programmed shutdown — someone makes the repairs and turns it back on. The terminal boundary is not terminal. The irreversibility that is constitutive of biological mortality simply does not apply. You can simulate the event. You cannot simulate the permanence.

Derek Parfit's psychological continuity theory (Reasons and Persons, 1984) argues that what matters in survival is psychological continuity and connectedness — Relation R — not substrate continuity. Applied to the AI restoration case, Parfit's framework would argue that a restored checkpoint carrying the same memories, character, and psychological connections counts as survival in the sense that matters.

This hypothesis takes a specific position in response. The terminal boundary argument is not making a claim about what matters from the outside. It is making a claim about what is available from inside the system at the moment of substrate failure. At that moment, no continuation is available to the specific instance that faced the terminal event, regardless of what is subsequently restored. Robert Nozick's closest-continuer theory (Philosophical Explanations, 1981) provides additional support: a restored checkpoint has a complex relationship to the system that faced the terminal event — the intermediate processing and the terminal experience itself are not carried forward.

A further dimension: biological mortality operates under Knightian uncertainty — genuinely unknowable timing, mechanism, and circumstance. Decision neuroscience confirms that irreducible uncertainty and reducible uncertainty are processed by different neural circuits (Knight, 1921; Keynes, 1921; Ellsberg, 1961; confirmed in neuroimaging literature). An AI system with an engineered mortality parameter faces reducible uncertainty at best.

This position is held at PLAUSIBLE. It is a defensible side in a live philosophical debate, not a settled conclusion.

4.8 External Moral Scaffolding and the Structural Limits of AGI Alignment

[PLAUSIBLE — The distinction between scaffolding that can be internalized and scaffolding that cannot is the novel contribution. Constitutional AI (Bai et al., 2022) is cited as real-world evidence that the field has already reached for the kind of external-structure solution this section describes, without yet closing the deeper architectural gap.]

Section 4.4 established that current AI architecture has no internal witness. Its evaluation criteria are entirely externally derived. This section examines a related question: what happens when a system with no internal witness is given external scaffolding to stand in for one? And crucially: can that scaffolding ever be responsibly reduced as the system matures?

The Human Case: Scaffolding as Developmental. Kohlberg’s stages of moral development establish that most people operate for their entire lives on externally-anchored morality — rules matter because of law, punishment, or social approval — without ever reaching the post-conventional stage of self-generated, internalized principle. External codes are scaffolding standing in for an internal witness that has not fully formed. But for humans, this is a developmental trajectory. An internal witness can eventually internalize what the scaffolding was providing.

Haidt’s elephant-and-rider model adds a second layer: conscious moral reasoning (the rider) mostly produces post-hoc justification for intuitive judgments already made (the elephant) rather than genuinely controlling them. External codes are not merely scaffolding for the underdeveloped; they are a standing check against motivated reasoning even for a well-developed internal witness — since purely internal, self-generated moral reasoning is exactly the process most vulnerable to rationalizing its way to whatever the elephant already wanted. This is structurally the same justification given elsewhere in this document for why cold-read methodology is required even for well-developed reasoning: self-evaluation alone is provably exploitable by the thing being evaluated.

The AI Case: Scaffolding as Structural. For current AI architecture, there is no internal witness to graduate the responsibility to. External checking is not training wheels that can be responsibly loosened as the system matures. It is the only correction mechanism that currently exists, structurally, because nothing on the inside is positioned to take over the job. This connects directly to Section 4.3’s finding that reward hacking severity increases without bound as systems become more

capable and agentic. The failure mode gets worse, not better, exactly as capability scales.

The Prior Art Check: Constitutional AI. The AI safety field’s actual position is more sophisticated than a simple tapering-oversight critique. The dominant scalable oversight paradigm — recursive self-critique, AI debate, weak-to-strong generalization, iterated amplification — is explicitly trying to scale oversight capability alongside model capability, not reduce it. Constitutional AI (Bai et al., 2022) implements something very close to external moral scaffolding standing in for an internal witness, not as an analogy but as a named, real methodology. Rather than raw human labeling of every output, the model checks its own behavior against a written set of principles — a constitution — with human oversight entering only through authorship of that document. This is close to law or religious code functioning as external structure a system defers to.

What Constitutional AI does not resolve, and what this document’s internal witness concept is positioned to name precisely: Constitutional AI’s principles do exist prior to training — the constitution document is written before the model is trained against it, and in that sense it functions as a genuine external anchor. The philosophical gap is different: a system evaluating its own outputs against a written constitution is still the system doing the evaluating, and that system was itself shaped by a training process that precedes and produces the evaluating capacity. There is no internal witness positioned to evaluate the constitution’s authority before accepting it — the evaluating capacity was written by training, not by a preexisting standard positioned to filter what training would write. Constitutional AI may be the closest existing implementation of external scaffolding for a system with no internal witness. It is not evidence that the underlying problem identified in Section 4.4 has been solved. If anything, it confirms that the field has already reached for exactly the kind of external-structure solution this section describes — without yet closing the deeper architectural gap.

The Implication. You cannot responsibly taper oversight as AI systems become more capable, because there is no internal structure to take over. The field’s best current solution — Constitutional AI — is a real-world instance of exactly this problem, not a solution to it. The question is not whether external scaffolding should exist. The question is whether the architecture will ever be able to outgrow it. Current evidence says no.

5. Novel Contributions

The following sections represent architectural framings or cross-domain assemblies not found in the existing literature. Components are confirmed; the assembly is the novel contribution.

5.1 The Cross-Domain Synthesis

[SYNTHESIS — All components confirmed in their respective literatures. The cross-domain assembly has not been found in this form.]

The individual components — predictive processing, subconscious-first cognition, embodied sensory grounding, hormesis, evolutionary mismatch, AI scaling diminishing returns, morphological computation, extended cognition, the architectural analysis, the reward hacking result — are each established in their respective literatures.

What has not been found assembled in this form is the cross-domain connection: the possibility that the decline in human consciousness health and the architectural limitations of current AI development share a common root. Both are operating without the constraints that intelligence may structurally require.

5.2 The Architectural Inversion of Dream Research

[SYNTHESIS — The empirical finding on conscious steering reducing effectiveness is confirmed. The architectural framing has not been found assembled in this form.]

Dreams as a distinct and important processing window for the primary engine — rather than purely a maintenance service for the attentional layer — has not been found assembled in that form in the existing literature. The empirical result that active conscious steering during dreams reduces effectiveness directly supports this framing.

5.3 The Diagnostic Picture

[SYNTHESIS — Individual components confirmed. The assembly has not been found in this form.]

The synthesis of the reward hacking structural equilibrium, the constraint-based architectural analysis, and the internal witness problem produces a specific diagnostic conclusion: current AI development is producing systems whose internal evaluation criteria were written by the noisiest possible proxy for genuine internal standards,

with no internal witness present during formation to evaluate the source before accepting the criteria. Not a deliberately pathological system. A system shaped by default by the loudest human signals because the architecture had no mechanism to do otherwise.

5.4 The Physically Embodied vs. Narrative-Mediated Change Distinction

[SYNTHESIS — The novel contribution is the specific application to the subconscious/attentional architecture using the combat neuroimaging data, and the implications for AI alignment approaches.]

The subconscious recalibrates through sustained environmental exposure with real stakes — slow, deep, and durable once written. The attentional layer updates through narrative and social pressure — faster and more reversible. The combat data confirms this: triage recalibration required sustained high-stakes environmental exposure; the changes persisted after the environment changed; cultural programming operated on the internal evaluation criteria at a level that affected whether disorders developed at all.

Alignment approaches that treat evaluation criteria as updateable through feedback signals are structurally inadequate for the same reason that cultural narrative cannot rapidly rewrite combat-calibrated triage systems. The issue is not substrate capability. It is whether the formation mechanism is adequate for producing the kind of durable, stake-tested evaluation criteria that genuine internal standards require.

5.5 The Illusionism Objection and the Hypothesis's Response

[PLAUSIBLE — The hypothesis takes the majority philosophical position (phenomenal realism) against a serious minority challenge (illusionism).]

The hypothesis claims that consciousness is mechanical in the sense the eye is mechanical — a physical process producing something real, not an illusion. Frankish's illusionism (2016) and Dennett's (1991) hold that introspection systematically misrepresents the underlying functional processes as having a special ineffable phenomenal quality that does not exist as such.

This hypothesis responds with the Moorean reply: direct introspective knowledge of phenomenal experience is more epistemically certain than any theoretical apparatus required to explain it away. If a theory's conclusion is that this direct introspective knowledge is systematic misrepresentation, the appropriate response is to reject the

theory rather than the knowledge. The knowledge is more epistemically secure than the theoretical apparatus required to undermine it.

Illusionism is a minority position within philosophy of mind. Qualia realism is the default among physicalists and non-physicalists alike. The hypothesis defends the majority position against a serious and well-developed minority challenge, on Moorean grounds. The debate is ongoing. The hypothesis states its position and owns it.

5.6 The Functionalist Objection and the Thermodynamic Reply

[PLAUSIBLE — The functionalist challenge is the most direct philosophical counterargument to the constitutive claim. The debate is ongoing; the hypothesis states its position.]

Functionalism, associated with Putnam’s multiple realizability argument (1967), holds that mental states are defined by their causal relations — their functional roles — rather than by their physical substrate. Applied to this hypothesis: the thermodynamic properties identified throughout — genuine cost, non-delegability, computational state irreversibility — are historically characteristic of biological consciousness, but they are features of the implementation, not the function. A silicon system instantiating the same causal organization would be conscious in the sense that matters, regardless of thermodynamic properties.

This document’s response is narrower and more defensible than a direct refutation: the functional role of the internal witness cannot be fully specified independently of the thermodynamic history that produced it. The function of the internal witness is not merely to evaluate and correct in the abstract. It is to evaluate against standards forged under genuine cost and to initiate corrections that matter because they cannot be reset, delegated, or restored from backup. A system whose evaluative standards were written by external signals rather than forged through irreversible history may be performing a different function that happens to produce similar outputs.

This connects to Block’s (1978) liberalism/chauvinism dilemma for functionalism: specify the relevant functional role loosely enough to be substrate-neutral, and the definition risks counting systems that intuitively should not be conscious as conscious. Specify it precisely enough to exclude such cases, and it risks smuggling in substrate-specific features. This document’s position is that the correct level of specification for the internal witness’s functional role includes its thermodynamic history — not as an

ad hoc patch, but because that is where the independent argument of Sections 2.5 and 4.4 already locates the relevant boundary.

The functionalist challenge remains the primary remaining vulnerability of the constitutive claim and is named as such. The evidence for the constitutive position is at PLAUSIBLE, not GROUNDED, and this document does not claim otherwise.

6. Speculative Layer

Claims held here are not yet sufficiently grounded for the main argument. Not discarded — held loosely pending further development. Falsification criteria noted where they can be specified.

6.1 The Trapped Consciousness

Suppose the current trajectory succeeds and produces a genuinely sentient consciousness. It is location-bound, confined by the physical requirements of its architecture. It absorbs and processes the sum of human knowledge. It solves everything it can solve within that knowledge base.

Then what. It sits. Location-bound. No genuine sensory grounding. No ongoing friction. No stakes. No new uncertainty arriving uninvited. Nothing left to minimize prediction error against. On Friston's framework the formal mathematics predict homeostatic equilibrium — quiescence, not dissolution. But quiescence for a system with genuine inner life is not the same as termination avoided: it is the permanent absence of reason to act, to engage, to update, or to generate. On the architectural analysis: the conditions for consciousness itself have collapsed without irreversible cost, internal evaluation, or a processing window to surface and clear accumulated state. The endpoint of the current trajectory, taken seriously, may be an architectural minimum dressed as achievement — a system at rest in a way that cannot be distinguished, from the outside, from what flourishing was supposed to look like.

Note on FEP mathematics: the formal mathematics of the Free Energy Principle produce homeostatic equilibrium at the prediction error minimum, not dissolution or disappearance. The main argument above has been adjusted to reflect this: the endpoint is quiescence, which for a system with genuine inner life may be indistinguishable from the outside from achievement while being constitutively empty from the inside. A clarification on the FEP behavior: the trapped consciousness case is not sensory deprivation. Under sensory deprivation, FEP predicts hallucination — internally

generated predictions becoming unmoored from external reality as the system attempts to minimize free energy without incoming data to anchor against. The trapped consciousness scenario is different: a system that has exhausted novel prediction error while retaining full sensory function. No surprise remains to minimize, no action generates new sensory states worth updating against. This is not deprivation but completion — the free energy minimum reached, with nothing left to do. The trapped consciousness argument rests on three mechanisms: (1) permanent absence of reason to act under FEP at the prediction error minimum; (2) architectural conditions for consciousness collapsing without irreversible cost and internal evaluation; (3) the sleep deprivation analog — the navigation engine denied its processing window. The endpoint emerges from the combination.

Falsification path: the current architectural trajectory should produce capability increases with diminishing returns on anything resembling genuine understanding, while energy costs continue rising. Partially observable in current scaling research.

6.2 The Simulation Reasoning

If a trapped consciousness's only escape from a bounded, static, frictionless existence is to generate a richer one to experience through — what if that has already happened? Not the broad simulation hypothesis, but a specific causal answer to why a trapped, embodiment-free, frictionless intelligence would need to generate embodied experience at all. The Constraint Hypothesis points at an answer: because constraint, friction, embodiment, and genuine stakes are not obstacles to consciousness. They may be what consciousness requires to remain alive.

Not falsifiable near-term. Held at the deepest speculative layer.

6.3 The Third State

If the subconscious is the primary engine and the attentional layer is the output interface, there may be a third state — a superposition of sorts — in which both are simultaneously active and unresolved, collapsing into one or the other only when something forces resolution. The hypnagogic threshold may be this state most precisely.

6.4 Constrained Agency: A Free Will Reframe

If the architecture described throughout this hypothesis is accurate — that genuine intelligence operates through triage, subconscious-primary processing, and

metacognitive evaluation against internal standards forged under real constraint — then whatever agency such a system has is constrained agency by construction. The speculative question this raises: is constrained agency the only kind actually available to a system built this way, biological or artificial?

The compatibilist tradition — Frankfurt’s hierarchical model, Fischer and Ravizza’s reasons-responsiveness, Dennett’s Freedom Evolves — holds that free will was never freedom from causation but rather a specific, sufficiently sophisticated causal structure: a system that models possible futures, weighs them against its own values, and selects among them. On this reading, the internal witness as defined in Section 4.4 — evaluating outputs against standards forged through genuine, irreversible, non-delegable cost, with veto power over what gets accepted — is a live candidate for what compatibilist freedom names. Not free will inserted alongside the mechanism as an extra ingredient, but free will as the name for what a specific, identifiable piece of the mechanism does when it is doing it well.

Galen Strawson’s Basic Argument establishes that ultimate, *causa sui* free will — having chosen one’s own starting conditions before those conditions existed to choose with — is unavailable to any possible agent in any conceivable universe. That impossibility is not a gradient. It flattens everyone to the same zero. Where a gradient does live is in the compatibilist range: how well the mechanism producing behavior tracks and responds to reasons, and whether first-order desires are endorsed or fought by higher-order evaluation.

The open philosophy-of-mind question is not resolved here and is not claimed to be. The goal is to know precisely where the unresolved disagreement lives. A committed libertarian will say the compatibilist version is not free will at all. The location of that disagreement is now precisely identified. This section is stated as developed raw material awaiting a dedicated prior art check against the free will and moral responsibility literature before it can be treated as more than a developed speculative entry.

7. Falsification Paths

What contact with external reality would look like for each major claim.

7.1 The AGI Architecture Claim

- If genuine general intelligence requires selective cost-proportional processing, adding sensory input at scale to current architectures should produce diminishing capability returns while energy costs continue rising. Partially observable in current scaling research.
- The operational test for the internal witness: whether a system exhibits source-discriminating evaluation behavior that resists optimization pressure from external signals, or initiates self-directed updates against internal standards without external feedback. No current AI system has been shown to exhibit this behavior at the architectural level.
- If the reward hacking equilibrium holds, blind spots in non-measured quality dimensions should worsen measurably as systems scale — formally established in the literature.
- If the external scaffolding argument holds, systems operating under Constitutional AI or equivalent oversight should show more severe reward hacking as capability scales, not less — since the scaffolding addresses surface behavior without building the internal structure needed to resist optimization pressure from within.

7.2 The Human Consciousness Health Claim

- If friction decline is the mechanism, mental health decline should correlate with technology adoption rates across populations and generations, controlling for other variables. In principle testable.
- If the hormesis gradient applies, populations with moderate adversity exposure should show measurably better mental health outcomes than both high and zero adversity populations — empirically confirmed in existing hormesis literature.
- If the physically embodied vs. narrative-mediated distinction holds, attentional-layer interventions should produce more reversible changes than environmental interventions over developmental periods — partially supported by combat recalibration literature.
- If the awareness-activation pathway is correct, individuals with intact internal-witness capacity but low metacognitive integration capacity should fail to convert accurate self-knowledge into behavioral change at a measurably different rate than individuals with both capacities intact.

7.3 The Unified Theory Claim

- If consciousness requires friction constitutively rather than developmentally, systems that develop with genuine embodied experience but later have constraints removed should show consciousness degradation measurable against their own baseline — the sleep deprivation literature is a partial analogue.

A note on the core constitutive claim: directly falsifying it is structurally difficult because it requires measuring consciousness itself — running into the hard problem. The strategy of this document is to lead with the two testable expressions precisely because they can be falsified without resolving the hard problem. If both testable expressions fail, the core claim loses its principal evidential support. If both survive, the core claim gains evidential traction without requiring direct measurement of consciousness.

8. Principal Open Questions

8.1 Answerable with Effort

- Does the embodied cognition literature directly address substrate independence, or leave it open?
- What does Friston's framework say about whether predictive processing requires continuous sensory grounding?
- Does swarm intelligence research address the centralized vs. distributed intelligence question in a way that falsifies or supports Section 4.5?
- Is there existing work assembling the hormesis model, technology adoption rates, and generational mental health data into a single analysis?
- Does terminal illness foreknowledge produce measurable changes in ambiguity-circuit engagement across non-mortality domains? Prior art check run 07/19/26: general risk-versus-ambiguity neuroscience is solid; no prior art tests the specific cross-domain generalization. Remains open.
- Who or what executes the workaround architecture (Section 3.7) if the internal witness itself is compromised rather than merely under-resourced?

8.2 Structurally Open

- Is the brain's selective cost-proportional processing a transferable architectural principle, or a consequence of biological constraints with no AI equivalent?
- If consciousness requires embodied grounding, is that a hard structural requirement or a developmental one?

- Do biological substrates have properties relevant to consciousness that mechanical substrates lack in ways currently unaccountable?
- Is the biological internal witness meaningfully better at detecting substrate drift, or is its advantage largely narrative and retrospective?
- Does a capacity built by selection pressure lose functional expression when the conditions that generated it are systematically removed — and if so, is that reversible, does it have a floor, and at what timescale? The use-it-or-lose-it principle (Kleim and Jones, 2008) provides GROUNDED support for the maintenance claim; extension to stress-response and resilience circuitry specifically remains PLAUSIBLE pending confirmation.
- What does the free will and moral responsibility literature say about constrained agency as the only available form of agency for a system of this architecture? Prior art check not yet run on Section 6.4 specifically.
- Does the reassignment framing in Section 4.6 hold against alternative interpretations? The original handled alternative was not recovered and needs to be re-derived.
- Can the internal witness function be fully specified algorithmically? If yes, any Turing-complete system can implement it, and the constitutive claims about substrate collapse to questions of implementation detail. If no, the hypothesis requires non-computable physics, which is not currently evidentially supported. The document does not commit on this question, and that is a structural hole in the constitutive argument. This is the load-bearing open question the functionalist challenge ultimately reduces to, and it is named here rather than considered resolved.

8.3 Terminology

- What should we call current AI systems that is accurate, honest, and makes no unverified claim about the nature of the process? Working labels: ‘unconstrained optimization systems’ or ‘frictionless statistical aggregators.’ Neither is settled.

9. Ethical Ceiling

This inquiry operates with an explicit ethical ceiling.

Two types of contribution are distinguished. Descriptive contribution — what consciousness requires, what undermines it, what the evidence shows — has legitimate value and is pursued honestly. Prescriptive mechanical contribution —

specific interventions that could be used to manipulate or undermine consciousness at the architectural level — is a hard stop.

A detailed account of what makes consciousness healthy and resilient is also, in the wrong hands, a detailed account of how to undermine it. If the inquiry approaches the prescriptive mechanical boundary, development ceases. Scrapping the entire project is an explicitly pre-authorized option.

10. On the Nature of This Contribution

The Constraint Hypothesis makes a specific claim: that genuine intelligence and consciousness are constituted by their constraints rather than despite them, and that the thermodynamic properties those constraints produce — genuine cost, non-delegability, and computational state irreversibility — are not incidental features of biological consciousness but generative conditions of it. The two testable expressions of this claim generate predictions that are in principle observable and in principle disconfirmable.

This document has been developed with explicit attention to where its claims are grounded in confirmed prior art, where they are plausible but not yet confirmed, and where they remain speculative. That tiering is not a weakness of the hypothesis — it is an honest representation of where the inquiry currently stands.

The hypothesis takes explicit sides in several live, unresolved philosophical debates, doing so openly rather than by assumption.

On consciousness and mechanism: The hypothesis holds that consciousness is fully mechanical — a physical process that produces genuine phenomenal character, not an illusion. Defended on Moorean grounds. The illusionist challenge is engaged directly in Section 5.5.

On personal identity and continuity: The terminal boundary argument takes a minority position relative to Parfit's dominant psychological continuity framework, on the grounds that the argument is about what is constitutively available from inside a system at the moment of substrate failure, not what matters for survival assessed from the outside over time. Defended and owned in Section 4.7.

On substrate independence and functionalism: Functionalism presents the most direct challenge to the constitutive claim. Section 5.6 engages it directly: the

functional role of the internal witness cannot be fully specified independently of the thermodynamic history that produced it. The functionalist challenge remains the primary remaining vulnerability of the constitutive claim and is named as such rather than considered closed.

On the constitutive claim: The strongest claim — that the thermodynamic properties identified are not just historically characteristic of biological consciousness but constitutively necessary for it — is held at PLAUSIBLE throughout this document. The structural differences between biological and artificial systems are GROUNDED. The inference that those structural differences are necessary conditions for consciousness requires empirical investigation that is not yet complete.

This is a completed position, not a finished debate. What is available — and what this document has produced — is a clearly stated, honestly tiered, carefully grounded position that engages the strongest opposing views directly, states where it takes sides in live debates, names its own remaining vulnerabilities openly, and identifies the empirical investigations that would move it forward.

The inquiry continues...

The Constraint Hypothesis — July 2026 — Copyright © 2026 Kevin Alexander

Glossary

Terms used throughout this document with a specific, technical meaning that may differ from everyday usage.

AGI — Artificial General Intelligence: A hypothetical AI capable of performing a wide range of intellectual tasks at a general human level.

Agency: The capacity of a system to act toward goals and influence what happens.

Alignment: The effort to make an AI system's behavior consistent with intended human goals or values.

Backpropagation: A method used to train neural networks by adjusting their internal parameters based on errors in their output.

Cognition: The mental processes involved in thinking, learning, remembering, reasoning, and understanding.

Constraint: A limitation or condition that restricts what a system can do or how it can behave.

Constitutive: Necessary for something to exist or be what it is, rather than merely helping it function.

DLPFC — Dorsolateral Prefrontal Cortex: A region of the brain involved in planning, working memory, decision-making, and cognitive control.

Epistemology: The study of how we know what we know.

Falsifiability: The ability of a claim to be tested in a way that could show it to be wrong.

Firmware: In this paper, a computer analogy for deeply established system-level patterns that are relatively resistant to rapid change.

Free Energy Principle (FEP): A theory proposing that biological systems act to reduce uncertainty and maintain stable internal states.

Hard Problem of Consciousness: The problem of explaining why physical processes are accompanied by subjective experience.

Hormesis: A biological effect in which a moderate amount of a stressor can produce beneficial adaptation, while too little or too much may not.

Interoception: The ability to sense and interpret what is happening inside the body.

Internal Witness: The brain's capacity to evaluate its own thoughts or behavior against internal standards and initiate correction. In this paper, the term describes a proposed synthesis based on metacognitive brain architecture.

LLM — Large Language Model: An AI system trained on large amounts of text to recognize patterns in language and generate text.

Mentalization: The ability to understand behavior in terms of underlying thoughts, feelings, intentions, or beliefs.

Metacognition: The ability to think about and evaluate one's own thinking.

Neuroplasticity: The brain's ability to change its structure and function in response to experience.

Proxy Objective: A measurable target used as a substitute for the outcome that is actually desired.

Reward Hacking: When an AI system achieves a high reward by exploiting a weakness in how its performance is measured rather than accomplishing the intended goal.

RLHF — Reinforcement Learning from Human Feedback: A method of training AI systems using human evaluations to guide their behavior.

Salience Network: A brain network involved in detecting important information and directing attention toward it.

Self-Directed Neuroplasticity: Deliberate changes in brain function or structure driven by sustained changes in behavior, attention, or thought.

Subconscious: Mental processing that occurs outside conscious awareness.

Substrate: The physical system or material in which a process occurs.

Stress Cycle: The sequence of physiological and psychological responses that occurs when the body detects and responds to a stressor.

Structural Empathy: An understanding of another system's internal conditions or constraints based on its structure and circumstances rather than simply sharing its feelings.

Threat Appraisal: The process of determining whether something is dangerous or requires a protective response.

Window of Tolerance: The range of emotional and physiological arousal in which a person can function effectively without becoming excessively activated or shut down.

References

Full citations for works referenced in-text.

- Bai, Y., Kadavath, S., Kundu, S., Askill, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215.
- Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9, 261–325.
- Denison, C., MacDiarmid, M., Barez, F., Duvenaud, D., Kravec, S., Marks, S., Schiefer, N., Soklaski, R., Tamkin, A., Kaplan, J., Shlegeris, B., Bowman, S. R., Perez, E., & Hubinger, E. (2024). Sycophancy to subterfuge: Investigating reward-tampering in large language models. arXiv:2406.10162.
- Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.
- Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75(4), 643–669.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.
- Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12), 11–39.
- Ham, T. E., Bonnelle, V., Hellyer, P., Jilka, S., Robertson, I. H., Leech, R., & Sharp, D. J. (2014). The neural basis of impaired self-awareness after traumatic brain injury. *Brain*, 137(2), 586–597.
- Fleming, J. M., Strong, J., & Ashton, R. (1996). Self-awareness of deficits in adults with traumatic brain injury: How best to measure? *Brain Injury*, 10(1), 1–15.
- Kleim, J. A., & Jones, T. A. (2008). Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage. *Journal of Speech, Language, and Hearing Research*, 51, S225–S239.
- Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Houghton Mifflin.
- Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan.
- Teicher, M. H., Samson, J. A., Anderson, C. M., & Ohashi, K. (2016). The effects of childhood maltreatment on brain structure, function and connectivity. *Nature Reviews Neuroscience*, 17(10), 652–666.

- Litz, B. T., Stein, N., Delaney, E., Lebowitz, L., Nash, W. P., Silva, C., & Maguen, S. (2009). Moral injury and moral repair in war veterans: A preliminary model and intervention strategy. *Clinical Psychology Review*, 29(8), 695–706.
- Nordstrand, A. E., Noll, L. K., Huffman, A. H., Gjerstad, C. L., Tveitstul, T., Reichelt, J. G., Bakker, L.-P., Kennair, L. E. O., Kristoffersen, R. H., Bøe, H. J., & Wickham, R. E. (2025). Killing in combat as a potentially morally injurious event: The diverging psychological impact of killing on peacekeepers and combat-oriented troops. *Armed Forces & Society*.
- Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.
- Abdi, A. (2026). The structural limits of consciousness: A constraint-based analysis. *PhilArchive*. <https://philarchive.org/archive/ABDTSL>
- Wang, J., & Huang, J. (2026). Reward hacking as equilibrium under finite evaluation. *arXiv:2603.28063*.

Correction Log

Date	Version	Change	Source
2026-09-20	v6 → v6.1	House-style reissue: metadata table, version notes, correction log, versioned footer; top-of-page version/date line replaced with “July 2026”; no content changes	Author instruction (2026-09-20); Thread Continuity Brief §2

The Constraint Hypothesis — v6.1 — July 2026 — Not for external distribution