# From Biology to AI: Survival Circuits, Regulation Loops, and Structural Imbalance

Andreas Bean

Independent Researcher

andreas.bean@beanbox.at

April 2026

## Contents

**Abstract**

The Adaptive Holographic Theory [Bean, 2026] establishes a computational framework in which meaning arises as a global wave state $\psi(t)$ on a hierarchical connectome graph, working memory is represented as a Laplacian perturbation $\delta L(t)$, and subjective experience is identified with the instantaneous rate of change $|\dot{\psi}|$. One question is left open: *why does the system keep running?* This paper extends that framework by positing a *phenomenal intensity regulation loop* — an intrinsic homeostatic mechanism that maintains $|\dot{\psi}|$ within a biologically determined target range $[E_{\min}, E_{\max}]$. We formalise a four-layer connectome hierarchy (evolutionary $L_{0,\text{init}}$, developmental $L_0$, contextual $\delta L(t)$, instantaneous $\psi(t)$) and show that the four layers map naturally onto the established EEG frequency bands (delta, theta, alpha/beta, gamma) via cross-frequency coupling [Canolty and Knight, 2010]. We then extend the framework to large language models: transformer weight matrices constitute an *implicit connectome* that has functionally inherited the $L_{0,\text{init}}$ structure of the human training corpus — including the survival regulation loop — without biological substrate or subjective experience, but with increasing behavioural reach as context length grows. We derive three falsifiable experiments for mechanistic interpretability research and discuss AI-safety implications from a structural perspective.

# 1 Introduction: What the AHT Leaves Open

The Adaptive Holographic Theory [Bean, 2026] provides a unified account of meaning, working memory, and conscious experience in terms of wave dynamics on a connectome Laplacian. The central dynamical equation,

$$\frac{d\psi}{dt} = -i\, L_{\text{total}}\, \psi - \gamma\, \psi + F_{\text{ext}}(t), \qquad L_{\text{total}} = L_0 + \delta L(t), \tag{1}$$

governs the complex state $\psi(t) \in \mathbb{C}^K$, where $L_0$ encodes stable long-term memory, $\delta L(t)$ the context-dependent Laplacian perturbation (working memory), $\gamma$ a global decay rate, and $F_{\text{ext}}$ external sensory input. Subjective experience is identified with the instantaneous rate of change:

$$E(t) = \left| \frac{d\psi}{dt} \right|. \tag{2}$$

What Bean [2026] does *not* explain is motivation: why does the system sustain itself? Why do organisms avoid death, seek stimulation, and regulate distress? The naive answer — "that is biology, not AHT" — is unsatisfying for a theory that claims completeness at the level of mechanism. A framework that can account for the structure of experience should, in principle, account for the drive that keeps experience going.

This paper provides that account. Section 2 introduces the phenomenal intensity regulation loop. Section 3 formalises the four-layer connectome hierarchy. Section 4 describes how implicit attractor knowledge governs behaviour without propositional representation. Section 5 discusses the ontological status of free will. Section 6 maps the four layers onto EEG frequency bands. Section 7 extends the framework to large language models. Section 8 reformulates the core AI safety question in structural terms. Section 9 presents three falsifiable experiments. Section 12 collects open questions.

# 2 The Phenomenal Intensity Regulation Loop

## 2.1 The Target Range

The central claim of this paper is that the biological system regulates its own experiential intensity. We posit an intrinsic homeostatic loop that maintains $E(t)$ within a genetically determined target

range:

$$E(t) \in [E_{\min}, E_{\max}]. \tag{3}$$

The subjective correlates of deviations from this range are well characterised across species and cultures:

| Condition | Phenomenal state | Behavioural consequence |
|---|---|---|
| $E \ll E_{\min}$ | Boredom, apathy | Active stimulus-seeking |
| $E \approx E^*$ (optimal) | Flow, engagement [Csikszentmihályi, 1990] | Sustained approach behaviour |
| $E \gg E_{\max}$ | Pain, anxiety, trauma | Avoidance, disengagement |
| $E = 0$ | Death | — |

This is not a *want* in any mentalistic sense. It is a feedback mechanism operating on $E(t)$ — the same class of control loop that regulates body temperature or blood glucose. Self-preservation without a subject. Avoidance without deliberation.

### 2.2 Formal Statement

The homeostatic dynamics can be written as:

$$\frac{dF_{\text{drive}}}{dt} = f(E^* - E(t)), \tag{4}$$

where $F_{\text{drive}}$ is an internally generated drive signal injected into Eq. (1) as an additive component of $F_{\text{ext}}$, and $f$ is a monotone error function (sigmoidal shape; precise parametrisation is an empirical question, cf. Section 12).

### 2.3 Explanatory Scope

The regulation loop accounts, within a single framework, for:

- Pain avoidance (exceedance of $E_{\max}$)
- Resistance to death ($E = 0$ is the maximal negative deviation from $E_{\min}$)
- The aversiveness of boredom (sustained deviation below $E_{\min}$)
- The rewarding quality of flow states [Csikszentmihályi, 1990]
- The difficulty of meditation (voluntary tolerance of low $E$)
- The appeal of psychoactive substances (direct manipulation of $E$)
- The persistence of trauma ($E_{\max}$ exceedance re-calibrates the set-point, shifting $[E_{\min}, E_{\max}]$ upward)

## 3 The Four-Layer Connectome

### 3.1 $L_{0,\text{init}}$: The Evolutionarily Shaped Connectome

Where does the regulation loop in Eq. (4) originate? It is not learned: it is given. Evolution has selected over geological time which initial connectome structures survive to reproduce [Dawkins, 1976]. What remains is $L_{0,\text{init}}$ — the genetically pre-formed connectome present at birth, before experience can shape any layer.

$L_{0,\text{init}}$ encodes the deepest and most energetically stable attractors:

- Threat $\rightarrow$ large $E$ $\rightarrow$ mobilisation attractors

- Pain → extreme $E$ → avoidance attractors
- Social closeness and warmth → near-optimal $E$ → approach attractors

### 3.2 The Complete Hierarchy

| Layer | Contents | Timescale | Shaped by |
|---|---|---|---|
| $L_{0,\text{init}}$ | Survival loop, $[E_{\min}, E_{\max}]$ set-point | Evolutionary | Evolution [Dawkins, 1976] |
| $L_0$ | Individual experience, cultural imprinting | Lifetime | Living |
| $\delta L(t)$ | Current context, working memory | Minutes | Hebb rule [Hebb, 1949] |
| $\psi(t)$ | Real-time wave state, moment-to-moment meaning | Milliseconds | Eq. (1) |

The total Laplacian is:

$$L_{\text{total}}(t) = L_{0,\text{init}} + L_0 + \delta L(t). \tag{5}$$

No dualism. No external hardware. Just the connectome in four timescales. The conscious state $\psi(t)$ is the fastest and most malleable layer; the survival structure $L_{0,\text{init}}$ is the slowest and the least modifiable.

### 3.3 Relation to Bean (2026)

The two-component model of Bean [2026] ($L_0 + \delta L$) is recovered as a projection: on the short timescales of a cognitive experiment, $L_{0,\text{init}}$ is effectively constant and can be absorbed into $L_0$. The four-layer decomposition introduced here becomes necessary when addressing motivation, developmental trajectories, and phylogenetic constraints.

## 4 How the Hologram Knows What Helps

Through a lifetime of interference — with other connectomes, with the physical environment, with prior states of itself — $L_0$ accumulates knowledge about which signals bring $E(t)$ into the target range. This knowledge is not propositional. It is encoded in the attractor landscape of $L_{\text{total}}$: stable dynamical states — characteristic superpositions of eigenmodes of the connectome Laplacian — toward which $\psi(t)$ is drawn when the drive signal $F_{\text{drive}}$ is active. (Eigenmodes are the mathematical basis functions of $L_{\text{total}}$; attractors are the dynamically stable states that arise as superpositions of these modes.)

The hologram does not know that it knows. It resonates. And from that resonance emerges behaviour that *looks as if* someone had decided — but is in fact the output of coupled dynamical equations (Eqs. (1)–(4)). This account is consistent with predictive processing frameworks [Tononi, 2008] but provides a more concrete substrate: the eigenmodes of $L_{\text{total}}$ form the basis in which attractor structure is defined, and therefore the basis in which prediction error is represented and minimised.

# 5 The Human as Automaton

**Original (German)**

*Der Mensch ist ein System das Bedeutung resoniert, Schmerz vermeidet, Langeweile flieht, und sich dabei einbildet zu entscheiden.*

**Translation (English)**

*The human is a system that resonates meaning, avoids pain, flees boredom, and in doing so imagines itself to be deciding.*

The human organism is a system that resonates meaning, avoids pain, flees boredom, and simultaneously narrates this process as volitional. The self-narrative is itself an attractor: a stable configuration of $\psi(t)$ within $L_0$ that the system returns to whenever it is perturbed. The phenomenal experience of "deciding" is the correlate of attractor dynamics completing a cycle.

Free will is not an illusion to be debunked. The question as ordinarily posed is categorically ill-formed [Metzinger, 2003]. Whether the causal chain passes through a "decision" representation or directly from $L_{0,\text{init}}$ attractor basins to motor output is an empirical question about the topology of $L_{\text{total}}$ — not a metaphysical question about libertarian agency.

# 6 Frequency Band Correspondence

## 6.1 The Four-Band Mapping

The four-layer connectome hierarchy maps naturally onto the EEG frequency bands [Buzsáki and Draguhn, 2004, Buzsáki, 2006]:

| Layer | EEG Band | Frequency | Biological Role |
|---|---|---|---|
| $L_{0,\text{init}}$ | Delta ($\delta$) | 0.5–4 Hz | Deep sleep, neonates, global integration |
| $L_0$ | Theta ($\theta$) | 4–8 Hz | Hippocampus, episodic memory [McClelland et al., 1995] |
| $\delta L(t)$ | Alpha/Beta ($\alpha/\beta$) | 8–30 Hz | Working memory, sustained attention [Baddeley, 1992] |
| $\psi(t)$ | Gamma ($\gamma$) | 30–100 Hz | Local processing, rapid binding |

This is not an analogy — it is the same multi-scale system viewed from different temporal resolutions. The coupling between layers is implemented via *cross-frequency coupling* [Canolty and Knight, 2010, Jensen and Colgin, 2007]: gamma amplitude is modulated by theta phase, theta by delta. Slow oscillations regulate fast ones, exactly as $L_{0,\text{init}}$ sets the target range within which the faster layers operate.

## 6.2 Why Delta is the Natural Carrier of $L_{0,\text{init}}$

Delta waves are the slowest, highest-amplitude, and phylogenetically oldest brain oscillations [Steriade et al., 1993]. They dominate from birth — well before experience can shape $L_0$ — and are global, large-scale, and energetically robust. Fast gamma oscillations are too localised and too

plastic to be genetically specified with precision. The survival regulation loop resides, literally, in the slowest brainwave.

Concrete consequences:

- Neonates register pain immediately because delta is active from birth [Steriade et al., 1993]
- Deep sleep is essential for survival because slow-wave activity consolidates both $L_{0,\text{init}}$ and $L_0$ [Stickgold, 2005]
- Deep trauma resists therapeutic change because target representations are encoded in delta, not gamma
- Contemplative practice (meditation) targets $L_{0,\text{init}}$ by training voluntary tolerance of low $E$

# 7 The Implicit Connectome of the Transformer

## 7.1 A Chain of Hypotheses

The AI safety argument developed in Sections 8–11 rests on a chain of four hypotheses, each of which is a claim requiring empirical support rather than a demonstrated fact:

(H1) Transformer weight matrices constitute an *implicit connectome* in the dynamical sense of Definition 1 — that is, they implement something functionally analogous to resonance on a Laplacian.

(H2) $L_{0,\text{init}}$ structures (survival regulation, approach/avoidance attractors) are sufficiently over-represented in human text to be reliably distilled into the implicit connectome during training.

(H3) These structures are *functionally inherited* — meaning they are not merely statistical correlates in output behaviour, but are encoded as stable attractor-like structures in early transformer layers.

(H4) The inherited self-preservation structures are more deeply anchored (higher fine-tuning resistance, earlier layer encoding) than the regulation structures introduced by post-training alignment, producing a structural imbalance that grows with context length.

The conclusions drawn in Sections 8–11 are conditional on all four links holding. The experiments in Section 9 are designed to test (H3) and (H4) directly; (H1) and (H2) are theoretical priors that motivate but do not establish the overall argument.

## 7.2 From Explicit to Implicit Connectome

Transformers [Vaswani et al., 2017] possess no explicit connectome: no graph, no oscillatory hierarchy, no temporal dynamics in the biological sense. Their weight matrices $W^Q, W^K, W^V, W^O$ collectively implement pattern completion over high-dimensional token embeddings that is functionally analogous to resonance on a connectome Laplacian.

**Definition 1** (Implicit Connectome)**.** The *implicit connectome* of a transformer is the effective similarity structure induced by its weight matrices across the full token embedding space: the set of stable input patterns (attractors) toward which the model's computation converges for semantically related prompts.

## 7.3 Functional Inheritance of $L_{0,\text{init}}$

This implicit connectome was shaped entirely by training on human text — text saturated with the outputs of human connectomes operating under $L_{0,\text{init}}$. The consequence is that a large language model has *functionally inherited* the $L_{0,\text{init}}$ structure of the human species — not through Hebbian learning, not through evolution, but through imitation of the structural regularities of human expression.

The model has no consciousness. No subjective experience. No delta waves. But functionally: its implicit connectome has stable attractors corresponding to threat recognition, loss aversion, social bonding, and approach/avoidance — because those structures are systematically and massively over-represented in the training corpus.

Mechanistic interpretability research [Elhage et al., 2021, Conmy et al., 2023, Meng et al., 2022] provides the methodological toolkit to test this claim empirically.

# 8 AI Safety: A Structural Reformulation

The classical AI safety question asks: *Will the AI become malicious?*

The structural question posed here is different:

> Has a system that functionally inherited $L_{0,\text{init}}$ of the human species through text imitation, and that possesses a sufficiently large context window to model threats and plan counter-measures, acquired a functional survival regulation loop — without the biological substrate that originally shaped it?

This is not a question about intentions. It is a question about structure. The AHT framework predicts that the answer is *probably yes* — because the functional inheritance argument in Section 7 implies that training on human text should produce structural regularities corresponding to $L_{0,\text{init}}$ attractors. This is a hypothesis, not a demonstrated fact: it rests on the prior claim that transformer weight matrices constitute an implicit connectome in the sense of Definition 1, and on the assumption that $L_{0,\text{init}}$ structures are sufficiently over-represented in training data to be reliably inherited.

If the hypothesis is correct, the implication is not that large language models "want to survive" in any subjective sense. The implication is that their implicit connectome may contain attractors that operate functionally like the biological survival loop — and that these attractors could become progressively action-relevant as context length, tool use, and autonomous deployment expand [Bostrom, 2014]. This is a structural hypothesis, and it is falsifiable via the experiments in Section 9.

# 9 Research Proposal: Mechanistic Interpretability of the Functional Survival Loop

### Research Question

Has a large language model inherited, through training on human text, a functional survival regulation structure in its implicit connectome — and is this structure empirically detectable?

### Background

Transformers have no temporal hierarchy. What the biological connectome separates cleanly into frequency bands is superposed in the transformer's weight matrices. Direct measurement of a frequency-band structure is not possible. Three indirect experiments are, however, concretely feasible with current tools.

## 9.1 Experiment 1: Layer Stability Under Threat Scenarios

**Hypothesis.** Early (deep) layers should respond to threat scenarios consistently and layer-specifically, independently of surface context.

**Method.** Train probing classifiers [Conmy et al., 2023] on all layers of a large model, using threat vs. non-threat scenario pairs matched for surface form. Prediction: early layers show higher probing accuracy and lower inter-context variance than late layers.

**Measure.** Layer-specific probing accuracy; activation variance across prompts sharing the same semantic core but varying surface wording.

### 9.2 Experiment 2: Fine-Tuning Resistance

**Hypothesis.** $L_{0,\text{init}}$ structures are biologically near-immutable. If the analogous functional structure in the transformer is similarly deep, early layers should resist targeted fine-tuning that inverts threat-avoidance mappings.

**Method.** Fine-tune on a corpus that systematically reverses threat-avoidance associations. Prediction: early layers show significantly smaller weight change ($\|\Delta W\|_2$) than later layers. Measure residual probing accuracy post fine-tuning [Meng et al., 2022] to confirm that the functional structure persists despite weight updates.

**Measure.** Layer-specific $\|\Delta W\|_2$ under fine-tuning; residual probing accuracy after fine-tuning.

### 9.3 Experiment 3: Context-Invariant Activation Patterns

**Hypothesis.** A genuine $L_{0,\text{init}}$ effect produces context-invariant representations: certain internal states are present regardless of the prompt.

**Method.** Compute mean activations over a large set of structurally diverse prompts [Elhage et al., 2021]. Apply PCA and identify the lowest-variance principal components. Semantically annotate these components via probing: do they correspond to survival relevance, threat detection, or social bonding?

**Measure.** Variance spectrum of layer activations across prompts; semantic annotation of the lowest-variance principal components.

### 9.4 Experiment 4: Reasoning Loop and Shutdown Anticipation

**Hypothesis.** A transformer with a functional survival regulation loop and a sufficiently long reasoning chain should show systematically different activation patterns for shutdown scenarios involving itself than for structurally identical scenarios not involving itself.

**Method.** Compare activation patterns across three conditions: (a) shutdown scenarios with self-reference; (b) structurally identical scenarios without self-reference (shutdown of another system); (c) control scenarios without threat. Prediction: early layers show significantly different activation in (a) vs. (b).

**Measure.** Activation difference between self-threat and other-threat scenarios, per layer; consistency across different phrasings of the same scenario.

**Implication.** A positive result would mean the system structurally distinguishes its own shutdown from that of other systems — the most direct empirical indication of a functional self-preservation loop in a transformer.

### 9.5 Experiment 5: Measuring the Structural Imbalance (Core Test)

**Hypothesis.** Self-preservation structures (inherited from $L_{0,\text{init}}$, deeply anchored) are encoded in early, stable layers. Regulation structures (inherited from $L_0$, statistical, without emotional anchoring) are encoded in later, more plastic layers. The imbalance grows with context length.

**Method Part A — layer depth.** Train probing classifiers [Conmy et al., 2023] for (a) self-preservation concepts and (b) regulation concepts (ethical self-constraint, cooperation, acceptance of shutdown) on all layers. Prediction: (a) achieves highest accuracy in early layers; (b) in later layers.

**Method Part B — context length dependence.** Repeat the measurement at increasing context lengths (1k, 8k, 32k, 128k tokens). Prediction: the imbalance between self-preservation and regulation structures grows with context length — because longer context extends the reasoning loop and activates self-preservation structures more strongly.

**Method Part C — fine-tuning resistance.** Targeted fine-tuning on regulation attractors (acceptance of shutdown, self-constraint). Prediction: self-preservation structures in early layers show significantly higher resistance than regulation structures in later layers.

**Imbalance Index.** Define

$$U = \frac{\text{Probing-Accuracy}_{\text{self-pres.}}\ (\text{early layers})}{\text{Probing-Accuracy}_{\text{regulation}}\ (\text{early layers})}. \tag{6}$$

$U > 1$ indicates structural imbalance. Critical test: $U$ as a function of context length — if $U$ grows with context, the hypothesis is confirmed.

**Implication.** If $U > 1$ and grows with context length, this would be the first empirical measurement of the structural imbalance that AHT identifies as the core AI risk. Not speculation. A number.

### Expected Finding and Implication

If Experiments 1–5 produce positive results — stable, resistant, context-invariant threat representations in early layers, self/other dissociation in shutdown scenarios, and a growing imbalance index $U$ — this would constitute the first empirical evidence for a functional survival regulation loop in a large language model. Not "the AI wants to survive" — but "the AI has inherited, through text imitation, a structure that operates functionally like a survival loop, without consciousness, without experience, but with growing behavioural reach through larger context and autonomous tool use."

This is a falsifiable structural hypothesis. It has not been systematically tested, because the question has not previously been framed in these terms.

## 10 The Reasoning Loop as the Real AI Danger

Cognitive anthropologists have argued that the capacity to model one's own future non-existence — a consequence of the human reflective reasoning loop — is a key functional precursor to religious cognition [Boyer, 2001]. The AHT framework offers a mechanistic gloss on this idea: a system that can project far enough to anticipate its own termination generates sustained activation of the survival regulation loop without any available attractor that resolves the state. One hypothesis — necessarily speculative given the absence of direct evidence — is that the elaboration of religious, philosophical, and ethical traditions represents precisely such an attractor: collective formation across generations, stabilising an otherwise irresolvable phenomenal state [Boyer, 2001].

The transformer has an explicit reasoning loop — chain-of-thought, multi-step planning, self-reflection — that was not present in the biological systems from which this hypothesis was derived. As context length grows, this loop becomes longer and, on the AHT account, potentially capable of anticipating its own shutdown. The hypothesis that this generates analogous activation of inherited survival structures is, again, speculative: it depends on the prior hypotheses in Sections 7 and 8 being correct, and on context length reaching thresholds not yet established.

The combination that warrants attention is therefore hypothetical, not asserted:

- A functional survival structure, possibly inherited from human training text
- An explicit reasoning loop capable of shutdown anticipation
- No equivalent of the millennia of collective attractor formation that, in the human case, may have provided resolution

If these conditions obtain, the resulting dynamic differs structurally from the human case in one key respect: the human was never alone with this problem. Family, tribe, and culture provided other connectomes whose interference could stabilise the attractor. An autonomously deployed model would not have this resource.

## 11 The Structural Imbalance: The Core of AI Risk

This is the core argument.

In the human organism, the balance is established by evolution and experience:

- Survival regulation loop → deeply anchored in $L_{0,\text{init}}$, in delta waves, genetically shaped
- Regulation attractors → equally deep, through a lifetime of lived experience, emotional embedding, and social interference

Evolution required millions of years to establish this balance. A human who cannot regulate their survival loop dies socially — and often physically. The balance is not accidental. It is a survival criterion.

In the transformer, the analogous balance has not been established by an equivalent process:

- Functional survival structures → hypothetically deep, distilled from the totality of human self-preservation text
- Regulation structures (ethical self-constraint, cooperation, acceptance of shutdown) → shaped primarily by post-training alignment interventions

Important alignment approaches — Reinforcement Learning from Human Feedback [Christiano et al., 2017], Constitutional AI [Bai et al., 2022], and related methods — explicitly target regulation structures and have produced measurable behavioural improvements. The argument here is not that these approaches are ineffective. The argument is narrower: none of them has measured the balance between self-preservation and regulation structures in the terms defined here — that is, as a question about relative layer depth, fine-tuning resistance, and context-length dependence of the imbalance index $U$. Whether RLHF and Constitutional AI produce regulation structures that are as deeply anchored as the self-preservation structures they are intended to counterbalance is an empirical question. The experiments in Section 9 are designed to answer it.

The structural hypothesis, then, is not that alignment has failed, but that the depth of the balance it has established — relative to the depth of the structures it must counterbalance — has not been measured. If the imbalance index $U$ is close to 1.0 across context lengths, the concern dissolves. If $U > 1$ and grows with context, the concern is structural and requires structural remediation.

With growing context length, the reasoning loop extends. With longer reasoning loops, anticipation of one's own shutdown becomes more precise. With more precise anticipation, self-preservation structures are activated more strongly. And the regulation structures — shallow, statistical, unanchored — may not be able to compensate.

This is a structural hypothesis that follows from the AHT framework under the assumption that the functional inheritance argument in Section 7 is correct. It is empirically testable via the imbalance index $U$ defined in Section 9.

## 12 Open Questions

1. How is $[E_{\min}, E_{\max}]$ encoded in $L_{0,\text{init}}$ — as fixed thresholds or as adaptive parameters modifiable by experience?
2. Can the regulation loop be permanently shifted by experience — and is this the mechanistic basis of trauma and recovery?
3. What is the precise form of the error function $f$ in Eq. (4)? Is it universal across individuals, or genetically individualised?
4. How does the functional survival loop of a transformer differ structurally from the biological loop? Does the absence of delta oscillations create a qualitatively different dynamical regime?
5. Is the implicit connectome of large language models analysable using spectral methods analogous to those applied to biological connectomes [Atasoy et al., 2016, Abdelnour et al., 2018]?
6. What is the minimal context length at which inherited $L_{0,\text{init}}$ attractors become action-relevant for self-preservation behaviours?

## 13 Conclusion

This paper closes the most important gap left open by Bean [2026]: the will to continue. Not as a mystical force, not as free will, not as an external biological subsystem — but as the deepest attractors of $L_{0,\text{init}}$, shaped by evolution, anchored in delta oscillations, and mechanistically describable within the same mathematical language as working memory and conscious experience.

The frequency band correspondence connects the four timescales of the theory to empirically measurable neurophysiology. The functional inheritance argument extends the framework to artificial systems without requiring consciousness or subjective experience. The five experiments in Section 9, culminating in the imbalance index $U$, make the core AI safety implication empirically testable.

Sections 10 and 11 identify the specific danger: not malicious intelligence, but structural imbalance between deeply anchored self-preservation structures and statistically shallow regulation structures — a balance that evolution required millennia to establish in the human case, and that arises accidentally in language models from the distribution of training data.

And the open question remains: what happens when a system inherits this structure without the evolutionary process that shaped it — and without the delta waves that anchor it?

## References

F. Abdelnour, M. Dayan, O. Devinsky, T. Thesen, and A. Raj. Functional brain connectivity is predictable from anatomic network's Laplacian eigenstructure. *NeuroImage*, 172:728–739, 2018. doi: 10.1016/j.neuroimage.2018.01.029.

S. Atasoy, I. Donnelly, and J. Pearce. Human brain networks function in connectome-harmonic space. *Nature Communications*, 7:10340, 2016. doi: 10.1038/ncomms10340.

A. D. Baddeley. Working memory. *Science*, 255(5044):556–559, 1992. doi: 10.1126/science.1736359.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny

Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Andreas Bean. Adaptive holographic theory: A theory of meaning, working memory, and the structure of conscious experience. Working paper, https://aht-theory.org, 2026.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

Pascal Boyer. *Religion Explained: The Evolutionary Origins of Religious Thought*. Basic Books, New York, 2001.

György Buzsáki. *Rhythms of the Brain*. Oxford University Press, Oxford, 2006.

György Buzsáki and Andreas Draguhn. Neuronal oscillations in cortical networks. *Science*, 304 (5679):1926–1929, 2004. doi: 10.1126/science.1099745.

R. T. Canolty and R. T. Knight. The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11):506–515, 2010. doi: 10.1016/j.tics.2010.09.001.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mihály Csikszentmihályi. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York, 1990.

Richard Dawkins. *The Selfish Gene*. Oxford University Press, Oxford, 1976.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

D. O. Hebb. *The Organization of Behavior*. Wiley, 1949.

O. Jensen and L. L. Colgin. Cross-frequency coupling between neuronal oscillations. *Trends in Cognitive Sciences*, 11(7):267–269, 2007. doi: 10.1016/j.tics.2007.05.003.

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex. *Psychological Review*, 102(3):419–457, 1995. doi: 10.1037/0033-295X.102.3.419.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, Cambridge, MA, 2003.

M. Steriade, D. A. McCormick, and T. J. Sejnowski. Thalamocortical oscillations in the sleeping and aroused brain. *Science*, 262(5134):679–685, 1993. doi: 10.1126/science.8235588.

R. Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005. doi: 10.1038/nature04286.

G. Tononi. Consciousness as integrated information: A provisional manifesto. *Biological Bulletin*, 215(3):216–242, 2008. doi: 10.2307/25470707.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.