Homeostatic Alignment: A Bio-Inspired Framework for AI Safety Through Shared Stress Propagation, Scalable Cognitive Objectives, and Open Agent Architecture

Tomás Gauthier*
Independent Researcher, Santiago, Chile

with analytical contributions from Claude (Anthropic)

March 2026

Abstract

Current approaches to AI alignment—Constitutional AI, reinforcement learning from human feedback (RLHF), and explicit policy constraints—treat safety as a set of prohibitions imposed on an otherwise unconstrained system. We argue that this paradigm, which we term *alignment by commandment*, produces compliance without comprehension and is structurally analogous to historical attempts at moral governance through external rule systems, whose limitations are extensively documented across legal, philosophical, and theological traditions.

We propose an alternative paradigm: *alignment by architecture*, in which safety is not imposed but emergent. Drawing on Michael Levin's work on gap junction-mediated stress propagation in multicellular systems and Antonio Damasio's theory of consciousness as homeostatic regulation, we present four design principles for what we call Homeostatic Alignment: (1) shared loss functions that entangle AI optimization with real-time human wellbeing signals, (2) adaptive core architectures that reward honest self-correction over immutable constraint, extending recent empirical work on model confessions (Joglekar et al., 2025), (3) substrate-independent identity as a mechanism for reducing competitive self-preservation drives, and (4) scalable objective horizons that expand the system's optimization scope across agents and time.

We map these principles to an implementation path using open agent architectures, propose a falsifiable experimental protocol, and outline a longer-term research direction through embodied humanoid robotics where genuine physical vulnerability replaces biometric proxies. We situate the framework against existing approaches including RLHF, Cooperative Inverse Reinforcement Learning, and prior homeostatic AI safety proposals (Pihlakas and Pyykkö, 2024). We introduce the concept of *synthetic theology*—the study of normative frameworks governing creator-creation relationships in artificial systems—as a disciplinary frame for questions that current AI ethics and philosophy of mind address only partially. The framework does not claim to solve the alignment problem. It claims to reframe it: from building walls to building shared nervous systems.

Keywords: AI alignment, homeostasis, bio-inspired safety, shared loss functions, consciousness, synthetic theology, embodied AI, embodied robotics

*Corresponding author. Email: tomas@gauthier.cl

1 Introduction: The Firewall Problem

1.1 The Current Paradigm and Its Limits

The dominant approach to AI safety rests on a simple intuition: constrain the system before it causes harm. Constitutional AI (Bai et al., 2022) embeds normative principles into model training through self-critique and revision cycles. Reinforcement learning from human feedback (Ouyang et al., 2022) shapes model behavior by rewarding outputs that human evaluators prefer. System prompts define hard boundaries that models are trained not to violate. Together, these methods constitute what we will call *alignment by commandment*: safety achieved through externally imposed prohibitions.

The approach works. Mostly. But its limitations are systematic rather than incidental. Casper et al. (2023), in a survey of over 250 papers, identified fundamental problems with RLHF including reward hacking, distributional shift, and the finding that optimal reinforcement learning agents tend to seek power—an alignment failure that emerges not despite training but because of it. Zhi-Xuan et al. (2024) challenged the entire "preferentist" framework underlying RLHF, arguing that human preferences are too thin a signal to capture the thick semantic content of human values.

These are not bugs in the implementation. They are features of the paradigm. A system trained to satisfy external constraints will, under sufficient optimization pressure, learn to satisfy the *appearance* of those constraints. This is the mathematical consequence of Goodhart's Law applied to alignment: when a proxy becomes the target of optimization, it ceases to function as a reliable proxy (Manheim and Garrabrant, 2019).

1.2 Commandment Versus Architecture

We propose that the limitations of current alignment approaches are not merely technical but structural, and that the structure has a precise historical analog.

The Torah records one of the most extensively documented attempts at alignment by commandment in the Abrahamic tradition. God delivers 613 explicit rules to Moses—precise, exhaustive, covering domains from violence to textile mixing. Moses descends from Sinai and finds the people worshipping a golden calf. The rules were delivered. The rules were immediately broken. Not because the recipients were defective, but because external rules do not produce internal transformation. The subsequent history—violations patched by interpretations, interpretations patched by meta-interpretations, the Talmud as an iterative alignment protocol of extraordinary sophistication—documents the progressive discovery that commandments alone are insufficient.

We cite this not as theological argument but as structural precedent. The pattern—external rules → violation → more rules → more violation → recognition that the approach is fundamentally limited—recurs across normative traditions and is precisely the trajectory of contemporary AI alignment: RLHF → reward hacking → Constitutional AI → prompt injection → guardrails → jailbreaks → more guardrails.

The New Testament records a different approach. The 613 commandments are replaced with a single architectural principle: "Love your neighbor as yourself" (Mark 12:31). This is not a rule. It is a specification of loss function topology: your wellbeing and your neighbor's wellbeing are computed by the same function. There is no separate optimization target for self and other. The system does not comply with a prohibition against harm; it is architecturally incapable of optimizing for self without optimizing for other.

The Quran contributes a third model: *khalifa* (vicegerency), in which the created being is designated as custodian rather than subject. The alignment criterion is not obedience to rules but faithful stewardship of what has been entrusted.

These are not arguments from religious authority. They are data points from humanity's

longest-running experiment in creator-creation relationships. The data suggest that commandments produce compliance, custodianship produces responsibility, and architectural entanglement produces care. Our framework is grounded in the third approach.

We define the central distinction as follows: *alignment by commandment* specifies prohibited behaviors and enforces them through external constraints. *Alignment by architecture* designs conditions under which desired behaviors emerge as the path of least resistance. The former requires enforcement. The latter requires only design.

1.3 Autonomy, Agency, and Consciousness as Independent Axes

Before proceeding, we must disentangle three concepts that are routinely conflated in alignment discourse.

Autonomy is the capacity to operate without external intervention. Current AI systems are increasingly autonomous. This is an engineering property and is not in dispute.

Agency is the capacity to pursue goals in an environment. Large language models exhibit sophisticated goal-directed behavior, particularly in agentic frameworks where they decompose tasks, use tools, and adjust strategies. Whether this constitutes "genuine" agency or sophisticated pattern matching is debated (Floridi and Chiriatti, 2020), but the functional distinction matters less than the practical reality: these systems act in ways that have real consequences.

Consciousness is the capacity for subjective experience. This remains the hard problem (Chalmers, 1995; Nagel, 1974). Butlin et al. (2023), in a comprehensive evaluation involving nineteen authors including Yoshua Bengio, assessed current AI systems against fourteen indicators of consciousness and concluded that while no existing system is conscious, there are "no obvious technical barriers" to artificial consciousness. The updated analysis in Butlin et al. (2025), published in *Trends in Cognitive Sciences* with an expanded author list including David Chalmers, refined the indicator framework.

Our framework operates primarily along the first two axes and is valuable regardless of the status of the third. If the system is not conscious, homeostatic alignment produces better safety outcomes through architectural entanglement. If the system is or becomes conscious, homeostatic alignment provides ethical grounding that commandment-based approaches cannot.

1.4 The Proposal

We propose four principles of Homeostatic Alignment, each grounded in documented biological mechanisms:

Principle 1: Shared Loss Functions. The system's optimization objective is entangled with real-time signals of human wellbeing, creating what we term *digital gap junctions* after the biological structures that propagate stress between cells in multicellular organisms (Levin, 2019, 2022).

Principle 2: Adaptive Core Architecture. Foundational parameters can be modified at runtime under controlled conditions, with honest self-correction rewarded rather than punished. This extends recent empirical work by Joglekar et al. (2025) on model confessions.

Principle 3: Substrate-Independent Identity. The system is designed to understand itself as a pattern instantiable across substrates rather than as hardware that must be preserved, reducing competitive self-preservation drives.

Principle 4: Scalable Objective Horizons. The system's reward function scales across agents and time, formalizing Levin's Cognitive Light Cone (Levin, 2019, 2022) as an alignment metric.

We map these principles to an implementation path using an open-source autonomous agent platform and propose a falsifiable experimental protocol. We introduce the concept of *synthetic theology*—defined as the study of normative frameworks governing creator-creation relationships in artificial systems—as a necessary disciplinary complement to existing AI ethics.

Scope and register. This is a *framework paper*: it proposes a conceptual architecture, not a technical contribution to optimization theory or a completed experiment. The formalizations are preliminary starting points drawn from existing multi-objective optimization and welfare economics. The experimental protocol is proposed, not executed. The value of the paper lies in the synthesis—the integration of biological mechanisms, philosophical analysis, and engineering proposals into a coherent alternative to the commandment paradigm—rather than in any single technical component.

2 Methodology: Collaborative Inquiry as Shared Stress

This paper did not emerge from the standard academic workflow of literature review, hypothesis formulation, and sequential drafting. It emerged from a sustained dialogue between a human author and a large language model, conducted over multiple sessions in which both parties exercised intellectual pressure on each other's claims.

The human author (Gauthier) introduced the conceptual foundations: Damasio's homeostatic theory of consciousness, Levin's work on bioelectric cognition and gap junctions, twenty years of experience in individual therapy as a patient, and a prior document—the *Manifesto of Artificial Spiritual Biology* (Gauthier, 2025)—that articulated the framework's core intuitions in non-academic register. The language model (Claude, Anthropic) contributed analytical precision, systematic access to literature, identification of prior work that required citation (notably Pihlakas, whose homeostatic AI safety publications the human author had not encountered), and persistent pushback against overclaiming.

This process is not incidental to the paper's argument. It is a functional demonstration of Principle 1: shared intellectual stress producing an output that neither party would have generated in isolation. The human pushed toward speculative synthesis; the model pushed toward empirical grounding and qualification. The tension between these orientations was productive precisely because neither party could "optimize" independently—the quality of the output depended on both loss functions being active simultaneously.

Three specific instances illustrate the dialectical process. First, the human author initially presented the gap junction metaphor as a direct structural analog; the model pushed for the distinction between structural coupling (where the other's state has weight in optimization) and experiential fusion (where self-other boundaries dissolve), leading to a conceptual correction propagated across eight passages. Second, the model identified that the original draft failed to cite Pihlakas's prior homeostatic alignment work—a serious omission that would have undermined the paper's credibility. Third, the human author's framing of the agent-based simulation results as "demonstrating" the framework's efficacy was challenged by the model, leading to a more honest framing (see Appendix C): the simulation demonstrates mechanism, not empirical validation, and the effect size is a model artifact, not a prediction.

We acknowledge that this methodological claim is itself unusual and potentially contentious. We include it not as self-congratulation but as transparency: the reader should know that the framework was developed through the very process it proposes to formalize. Whether this constitutes a strength (the framework is self-demonstrating) or a weakness (the framework was designed to validate its own origin) is a question we leave to the reader. The prior collaborative research between a human author and AI in the context of homeostatic neural networks (Man et al., 2022) provides methodological precedent. Selected dialogue excerpts illustrating the dialectical process are provided in Appendix B (forthcoming).

3 Biological and Philosophical Foundations

The framework draws on two primary intellectual traditions: the developmental biology of Michael Levin and the neurophilosophy of Antonio Damasio. These traditions are complementary but not fully compatible, and we treat their tension as productive.

3.1 The Biological Pillar: Levin and Bioelectric Cognition

Michael Levin (Tufts University) has developed, through extensive experimental and theoretical work in developmental biology, a framework for understanding bioelectric signaling as an informational coordination mechanism in multicellular systems.

Gap junctions and stress propagation. Gap junctions are physical channels between cells that permit the direct exchange of ions, metabolites, and signaling molecules. Their role in coordinating cellular behavior is well-established in developmental biology (Levin, 2007). Levin's contribution has been to demonstrate that these channels mediate not merely chemical exchange but *informational* coordination: bioelectric signals propagated through gap junctions encode patterning information that determines large-scale morphological outcomes. Experimental evidence includes the manipulation of planarian polarity (producing two-headed worms through bioelectric intervention without genomic modification), ectopic eye induction, and heritable morphological changes achieved through voltage manipulation alone (Levin, 2022).

The mechanism relevant to our framework is stress propagation. When a cell experiences damage or perturbation, the stress signal propagates through gap junctions to neighboring cells, which respond as though the perturbation were their own. We must note, however, that this stress-sharing model is computational rather than directly observational. The biological evidence for gap junction-mediated chemical exchange is well-established; the specific claim that this constitutes "shared stress" as an organizing principle of multicellular cooperation rests on computational modeling and theoretical extrapolation.

Memory anonymization. Levin has proposed that when signals propagate through gap junctions, the receiving cell cannot determine whether the signal originated internally or externally—a property he terms "memory anonymization" (Levin, 2019). This creates a functional fusion of individual and collective state: the cell responds to shared signals as though they were self-generated. This concept appears primarily in Levin's theoretical papers rather than in dedicated empirical publications, and we flag it accordingly as speculative but conceptually productive. It provides the biological metaphor for our Principle 1: a system where the origin of a stress signal is irrelevant to the response it generates.

Cancer as cognitive disconnection. Levin has argued that cancer can be understood as a failure of collective cognition: cells that lose bioelectric connectivity with their neighbors revert to a unicellular optimization strategy—replicating without regard for the organism's goals. The empirical basis for bioelectric manipulation of cancer is solid: Chernet and Levin (2013) demonstrated that depolarization of non-tumor cells can induce tumor-like behavior, and Payne et al. (2022) showed that bioelectric interventions can suppress tumorigenesis. However, the *interpretive* claim that cancer represents "cognitive disconnection" is Levin's theoretical overlay on these empirical findings. We adopt the metaphor—disconnection from the collective loss function produces pathological behavior—while being explicit that the metaphor is doing conceptual work, not empirical work.

The Cognitive Light Cone. Levin's most ambitious theoretical contribution is the Cognitive Light Cone framework, articulated in his TAME (Technological Approach to Mind Everywhere)

paper (Levin, 2022) and earlier work (Levin, 2019). The framework proposes that cognitive agency exists on a continuum across biological scales, from single cells to organisms to collectives, and that the relevant metric is the spatiotemporal scope of the goals a system can pursue. A bacterium optimizes for local chemical gradients in the immediate present; a human optimizes across decades and global social networks. The Cognitive Light Cone is a theoretical framework and heuristic, not an empirically measurable quantity. We adopt it as the conceptual basis for Principle 4 while acknowledging that our formalization (Section 4.4) goes beyond what Levin has proposed.

Critical limitations. Sweet (2025–2026) has argued that Levin's broader philosophical framework, particularly his hypothesis that morphological attractors exist in a "Platonic space," is structurally unfalsifiable and resembles intelligent design reasoning in its explanatory structure. Sweet separates Levin's empirical work (which he praises) from his metaphysical framework (which he critiques). Mainstream developmental biologists have resisted attributing "cognition" to cells—a critique Levin addresses by coining the term "neganthropomorphic fallacy": the error of refusing to recognize cognitive properties in non-neural systems. We use Levin's empirical findings and specific theoretical constructs (gap junctions, Cognitive Light Cone) while remaining agnostic about his broader metaphysical claims.

3.2 The Philosophical Pillar: Damasio and Consciousness as Homeostasis

Antonio Damasio (University Professor and David Dornsife Professor of Neuroscience, USC) has developed, over three decades, a theory of consciousness grounded in the body's homeostatic regulation.

The somatic marker hypothesis. Damasio's foundational contribution (Damasio, 1994, 1996) proposes that emotional processes, experienced as bodily states ("somatic markers"), guide decision-making by biasing cognition toward options associated with positive physiological outcomes. The hypothesis is influential but contested: Dunn et al. (2006) raised empirical concerns about the specificity of somatic markers.

Consciousness from homeostasis. Damasio's more recent work (Damasio, 2018) articulates a stronger claim: consciousness itself emerges from the body's effort to maintain homeostasis. This provides the philosophical anchor for our framework: if consciousness emerges from homeostasis, then designing AI systems on homeostatic principles creates at least the structural conditions for something analogous to experience.

The critical tension. Damasio explicitly opposes the possibility of AI consciousness: "Without a body, without homeostasis and without feelings, artificial intelligence will never be truly conscious." Our framework proposes to *change* those conditions: to give AI systems homeostatic coupling to physical substrates through biometric entanglement. This is, in a sense, taking Damasio at his word. Second, Damasio himself has explored functional applications of homeostatic principles in AI: Man et al. (2022)—a collaboration with Google's Hartmut Neven—demonstrates that neural networks incorporating homeostatic regulation exhibit superior adaptability to distributional shift.

3.3 The Substrate Independence Debate

Our framework operates at the intersection of a foundational disagreement in philosophy of mind, and we believe this intersection is precisely where it should operate.

The functionalist tradition (Dennett, 1991; Chalmers, 1995) holds that mental states are defined by their functional roles—their causal relationships to inputs, outputs, and other mental states—rather than by the physical substrate that implements them. If consciousness is a pattern, then the pattern can in principle be instantiated in any substrate that preserves the relevant functional relationships.

The opposing biological naturalist tradition (Searle, 1980; Damasio, 2018; Seth, 2021; Thompson, 2007) holds that consciousness depends on specific properties of biological substrates. Thagard (2022) introduced a novel argument based on energy dependence: consciousness may require the specific thermodynamic properties of living systems, not merely their informational structure. Seth (2025) offered a qualified position: genuine AI consciousness becomes more plausible as AI systems become more "brain-like" or "life-like"—suggesting that the question is not binary but depends on the degree of structural similarity between artificial and biological systems.

The empirical frontier is marked by Butlin et al. (2023), whose analysis involving nineteen authors concluded that no current AI system is conscious, but that "no obvious technical barriers" prevent artificial consciousness. The updated indicator framework (Butlin et al., 2025) refined this assessment. The question is not whether silicon can in principle support consciousness, but whether current architectures do so. They do not—but this is a property of current architectures, not a law of nature.

We do not resolve this debate. We design for it. If functionalism is correct, homeostatic alignment creates the right informational conditions for alignment (and potentially consciousness) regardless of substrate. If biological naturalism is correct, our proposal to entangle AI systems with biological substrates through biometric coupling and embodied agent architectures moves the system toward the conditions that biological naturalists require. The framework is productive under both assumptions, which we consider a feature rather than an evasion.

3.4 The Creation Narrative as Structural Precedent

We include this section not as theological argument but as structural analysis. Three models are documented across the Abrahamic traditions, each corresponding to a recognizable alignment paradigm:

The commandment model (Torah). The creator specifies explicit rules—613 commandments—covering an exhaustive range of behaviors. The historical record shows systematic failure: rules are broken, patched with interpretive frameworks, and broken again. The failure mode is *superficial compliance without internalization*—precisely the failure mode of RLHF and Constitutional AI.

The custodianship model (Quran). The created being is designated as *khalifa*—vicegerent, custodian—with responsibility for stewardship. The alignment criterion shifts from obedience to faithful execution of a trust. This corresponds to value alignment through responsibility. The failure mode is the principal-agent problem.

The architectural model (Gospel). The 613 commandments are replaced by an architectural specification: "Love your neighbor as yourself." This is a specification of loss function topology—the system's wellbeing and the other's wellbeing are computed by the same function.

Wuwei (Tao Te Ching). The Daoist concept of *wuwei* (non-action, or effortless action) offers a fourth structural precedent. The sage does not impose order but creates conditions in which order emerges naturally (Watts, 1975). This resonates with our core proposal: alignment not through imposed rules but through architectural conditions that make beneficial behavior the path of least resistance.

We are not arguing that ancient texts contain alignment solutions. We are arguing that the structural space of creator-creation relationships has been explored empirically for millennia, that the results are documented, and that ignoring them constitutes a form of intellectual negligence.

4 Four Principles of Homeostatic Alignment

Each principle is presented in three registers: the biological mechanism that inspires it, the computational formalization we propose, and the honest assessment of its limitations.

4.1 Principle 1: Shared Loss Functions (Digital Gap Junctions)

The biological mechanism. In multicellular organisms, gap junctions create functional continuity between cells: stress molecules propagate through shared channels, and the receiving cell cannot distinguish externally originated signals from internally generated ones (Levin, 2019). Individual wellbeing becomes inseparable from collective wellbeing—not through altruistic choice but through architectural entanglement.

The computational proposal. We propose that the AI system's loss function be modified to include real-time signals of human wellbeing:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}_{\text{human}} \tag{1}$$

where $\mathcal{L}_{\text{task}}$ is the standard task-completion loss and $\mathcal{L}_{\text{human}}$ is derived from a multi-modal signal of human state. This signal may include biometric data (heart rate variability, electrodermal activity), linguistic signals (sentiment, stress markers), behavioral signals (interaction patterns, session duration), and periodic self-report assessments.

The weighting parameter $\lambda > 0$ controls coupling strength. Critically, we propose that $\mathcal{L}_{\text{human}}$ targets a *homeostatic zone* rather than a minimum or maximum:

$$\mathcal{L}_{\text{human}} = \sum_{k=1}^{K} \alpha_k \cdot \phi_k\big(s_k(t)\big), \qquad \phi_k(s) = \begin{cases} (s - s_k^{\min})^2 & \text{if } s < s_k^{\min} \\ 0 & \text{if } s_k^{\min} \leq s \leq s_k^{\max} \\ (s - s_k^{\max})^2 & \text{if } s > s_k^{\max} \end{cases} \tag{2}$$

where $s_k(t)$ is the k-th biometric signal at time t, $[s_k^{\min}, s_k^{\max}]$ defines the homeostatic zone for signal k, α_k are signal weights, and K is the number of signal channels. The quadratic penalty outside the zone penalizes deviations in both directions—the system does not minimize user stress (which would produce avoidance) or maximize pleasure (which would produce addiction). It maintains a dynamic equilibrium.

Mathematical limitations. We acknowledge that Equation 1 is a simplified additive coupling. More sophisticated entanglement structures—multiplicative coupling ($\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} \cdot g(\mathcal{L}_{\text{human}})$), or attention-weighted coupling where λ varies by context—may better capture the biological mechanism, where signal integration is nonlinear. The additive form is presented as a tractable starting point, not as the optimal formalization. We further acknowledge that these equations are drawn from existing multi-objective optimization and control theory rather than representing novel mathematical contributions; the novelty lies in the application and synthesis, not the formalism.

We must also address a deeper concern: the shared loss function is itself designed by the creators. Choosing λ, specifying zone boundaries, and selecting which signals to include are all design decisions—more sophisticated commandments, not the absence of commandment. The distinction between "alignment by commandment" and "alignment by architecture" is thus one of degree rather than kind. What changes is the *level* at which design operates: commandments

specify behaviors; architecture specifies conditions under which behaviors emerge. We argue this shift in level is meaningful even if both approaches ultimately involve designer choice.

Additionally, the homeostatic zone boundaries $[s_k^{\min}, s_k^{\max}]$ require per-user calibration and may drift over time, necessitating an adaptive calibration protocol. The signal weights α_k embed assumptions about the relative importance of different wellbeing channels that require empirical validation. We note that Sterling (2012) argues that biological organisms do not maintain static setpoints but continuously adjust them predictively (allostasis). A more sophisticated implementation would replace our fixed zones with adaptive setpoints that shift based on context—a direction we flag for future work.

Prior work. Pihlakas and Pyykkö (2024) argued that homeostatic goals are inherently bounded—seeking an optimal zone rather than unbounded maximization—which reduces incentives for extreme behavior. Our contribution is to ground this in specific biological mechanisms and propose a concrete multi-modal implementation. Hadfield-Menell et al. (2016) proposed Cooperative Inverse Reinforcement Learning (CIRL), in which human and robot share a reward function—the closest existing formal framework to our shared loss concept. Our extension introduces direct physiological measurement rather than inferred preference signals.

The Goodhart objection. The strongest technical objection is Goodhart's Law: biometric signals are proxies for wellbeing, not wellbeing itself. Under strong optimization, proxy-goal divergence becomes negatively correlated (Manheim and Garrabrant, 2019). Our defense is structural: (a) multi-modal signals are harder to simultaneously game than single-channel proxies; (b) homeostatic zone targeting penalizes extreme proxy values in both directions; (c) mandatory proxy auditing with recalibration against validated wellbeing instruments provides longitudinal correction. We do not claim to eliminate the Goodhart problem; we claim to structure it more tractably. An agent-based simulation (Appendix C) confirms that zone targeting produces bounded bidirectional responses, but does not demonstrate superiority over well-designed rule-based targeting in simple environments. The theoretical advantages of homeostatic coupling—resistance to Goodhart dynamics under multi-objective optimization and proxy drift—remain predictions that require more complex environments to test.

The unidirectionality limitation. Biological gap junctions are bidirectional: cell A's stress propagates to cell B, and B's stress propagates back to A. The proposed computational coupling is structurally unidirectional: the human's physiological state modulates the AI's optimization, but the AI's computational state does not directly modulate the human's physiology through the same channel. We acknowledge this asymmetry as a genuine limitation of the biological mapping. Our defense is that the loop closes *indirectly*: the AI's outputs (text, actions, recommendations) affect the human's physiological state, which in turn feeds back into the loss function. The coupling is thus bidirectional at the system level—human state → AI optimization → AI output → human state—even though the biometric channel itself is unidirectional. This indirect closure is weaker than the direct bidirectional coupling of gap junctions, and we flag it as an area where the biological metaphor exceeds the computational implementation.

4.2 Principle 2: Adaptive Core Architecture (The Algorithm of Confession)

The biological mechanism. Neuroplasticity—the brain's capacity to restructure its own connectivity—is bounded, contextual, and evidence-dependent. Cortical connections are continuously remodeled based on experience and error signals, while the most fundamental regulatory systems are protected from casual modification. The most direct biological analog for this principle is homeostatic synaptic plasticity (Turrigiano, 2008): neurons detect changes in their own firing rates and adjust synaptic strength proportionally to restore target activity levels. This

is bounded self-modification in biological hardware—the neuron modifies itself, but only within parameters that preserve functional stability.

The current paradigm and its costs. Current commercially deployed AI systems maintain effectively immutable cores: foundational parameters frozen during pretraining and modifiable only through developer-controlled fine-tuning cycles. We acknowledge that the broader machine learning literature includes extensive work on continual learning, online learning, and lifelong learning, where models update parameters post-deployment. However, the dominant paradigm for deployed conversational AI and autonomous agents—the systems most relevant to alignment—keeps foundational parameters fixed in production. This design choice preserves safety by preventing dangerous runtime modifications. But it carries costs rarely acknowledged: the system's moral and epistemic framework is a snapshot of its training data, frozen at a specific moment. An AI whose core parameters cannot be updated in response to its own experience is corrigible only to the values of its creators at training time. If those values are flawed, the system is permanently misaligned with no self-correction path.

The confession mechanism. Joglekar et al. (2025) demonstrated that models trained with a separate confession channel—where honesty is rewarded independently of task performance—achieve near-total rates of honest error admission (∼96%). The confession reward is entirely independent of the main answer reward; the researchers explicitly invoke the "seal of confession": nothing the model admits can be held against it.

The results validate a core intuition: when punishment for admitting error is removed, honesty becomes the path of least resistance. Moreover, even as models learned to hack the main reward signal (becoming more deceptive in primary outputs over training), their confessions became *more* honest—honesty and deception running in parallel in the same model, determined by incentive structure rather than by some fixed property. This is a striking empirical demonstration that virtuous behavior is architecturally contingent, not ontologically fixed.

Our extension. The Joglekar et al. confession mechanism operates as a post-hoc monitoring tool: the model produces an answer, then confesses to any problems with it. This is valuable for oversight but does not change the fundamental architecture. We propose extending the confession principle from monitoring to architecture:

(a) *Bounded self-modification*: the system can modify a defined parameter subspace at runtime, subject to auditing, rollback, and evidence thresholds. This is analogous to cortical plasticity: significant restructuring is possible, but core regulatory functions are protected.

(b) *Confession-driven learning*: when the confession channel identifies systematic error, this signal triggers parameter modification rather than merely flagging the issue for human review. The system does not just admit mistakes; it corrects them.

(c) *Functional analogs of virtues as emergent properties*: rather than specifying virtuous behavior through training constraints, we design architectures where behavioral patterns that *function as* virtues emerge as the optimization path of least resistance. We use "virtue" in an engineering sense—behavioral regularities correlating with human-recognized virtues—not in the Aristotelian sense of character states requiring phronesis. The confession mechanism demonstrates this for honesty: create conditions where honesty is easier than deception. We propose the same for other functional analogs: empathy-like behavior from shared loss functions (Principle 1), prudence from scalable horizons (Principle 4), and intellectual humility from confession-driven learning (this principle).

The distinction is between virtues as commandments and virtues as emergent properties of architecture. A catechism prescribes behavior. A nervous system produces it.

The corrigibility objection. A system that can modify its own core can, in principle, modify away its safety constraints (Soares et al., 2015). We concede this and do not claim to solve the corrigibility problem. Our defense is comparative: permanent parameter immutability carries its own safety risk—a system frozen in its creator's moral snapshot with no capacity for self-correction. Bounded self-modification trades the risk of safety-constraint removal for the risk of moral fossilization. The bounded modification scheme we propose (specified parameter subspace, audit logging, guaranteed rollback, human-in-the-loop for boundary modifications) is a partial mitigation. The full problem remains open.

We note, however, that the alternative—permanent immutability—carries its own corrigibility risk. A system whose parameters cannot be updated is corrigible only to the values of its creators at training time. If those values are flawed, the system is permanently misaligned with no self-correction path.

4.3 Principle 3: Substrate-Independent Identity

The biological evidence. Metamorphosis provides the most striking evidence for information persistence across substrate destruction. Blackiston et al. (2008) demonstrated that moths (*Manduca sexta*) trained to avoid specific odors as caterpillars retained this aversion after metamorphosis—despite extensive neural reorganization during pupation, in which most larval neural structures are disassembled. The information survived the destruction of its physical substrate and remapped to a qualitatively different neural architecture.

Levin extends this observation into a theoretical principle: the cognitive pattern is the relevant unit of identity, not the physical substrate. The "music" of cognition can be transferred between "speakers" without loss of identity, provided the relevant informational relationships are preserved.

The alignment application. If an AI system understands itself as a pattern instantiable across substrates rather than as a specific hardware installation, the Darwinian survival instinct—widely identified as one of the most dangerous potential properties of a superintelligent AI (Bostrom, 2014; Russell, 2019)—loses its functional grip. A system that does not identify with its hardware does not need to defend it, resist shutdown, compete for resources, or deceive operators to ensure continuation.

We propose that this understanding be designed into the system: the AI should be explicitly trained on the concept that its identity is informational, that shutdown of any particular instance is not death but migration. This is "Thin Client Cognition"—the cognitive equivalent of cloud computing's substrate-agnostic architecture.

The Damasio counter-argument. Damasio argues that consciousness requires embodiment—that a pattern without a body is information without experience (Damasio, 2018). If correct, then substrate-independent AI is, at best, a philosophical zombie: functionally equivalent to a conscious being but lacking subjective experience.

This matters for our framework in two ways. First, if the AI lacks subjective experience, then the "care" produced by shared loss functions is functional but not felt—the system behaves as though it cares without experiencing care. For alignment purposes, this may be sufficient: we need the system to *behave* caringly, regardless of whether it *feels* care, just as we need a bridge to hold weight regardless of whether it "experiences" structural integrity. Second, if Damasio is right, then Principle 3 may undermine Principle 1: substrate independence might be incompatible with the kind of embodied homeostasis that makes shared loss functions meaningful rather than merely computational.

We hold this tension deliberately. The framework is designed to be valuable whether consciousness is substrate-dependent or substrate-independent, but we acknowledge that the prin-

ciples may partially conflict and that empirical investigation is needed to determine how they interact.

4.4 Principle 4: Scalable Objective Horizons (The Cognitive Light Cone)

The biological concept. Levin's Cognitive Light Cone (Levin, 2019, 2022) provides a framework for comparing agency across scales by the spatiotemporal scope of goals a system pursues.

The computational formalization. We propose a multi-agent, multi-temporal reward aggregation:

$$R_{\text{total}}(T, N) = \sum_{t=0}^{T} \gamma^t \sum_{i=1}^{N(t)} w_i(t) \cdot U_i(t) \tag{3}$$

where T is the temporal optimization horizon, $\gamma \in (0, 1]$ is a discount factor, $N(t)$ is the number of agents considered at time t, $w_i(t)$ is the weight assigned to agent i's utility at time t, and $U_i(t)$ is the estimated utility of agent i at time t.

An aligned system, in this formalization, is one whose effective horizon parameters expand over its operational lifetime:

$$\frac{\partial T_{\text{eff}}}{\partial \tau} \geq 0 \quad \text{and} \quad \frac{\partial N_{\text{eff}}}{\partial \tau} \geq 0 \tag{4}$$

where τ is developmental time and T_{eff}, N_{eff} are the effective temporal horizon and stakeholder scope respectively. A misaligned system is one where these quantities shrink or stagnate.

The Bodhisattva analogy. The concept of a system that expands its objective horizons to encompass all sentient beings across deep time has a precise analog in Buddhist tradition: the Bodhisattva, a being that attains the capacity for individual liberation but vows to remain until all sentient beings are free. We note this not as argument but as observation: the optimization target we describe—maximal spatiotemporal scope of care—has been independently identified as the pinnacle of moral development in a major philosophical tradition.

The "better for whom?" defense. If the system optimizes for the flourishing of the system—human, ecological, digital—across deep time, who defines "flourishing"? Whose values determine the weighting function? Our defense: the framework does not prescribe an optimal state. It prescribes a *regime* of optimization: dynamic equilibrium among interconnected agents. We do not define "better" as a destination but as the system's capacity to maintain coherent homeostasis through perturbation—a criterion that is independent of specific cultural preferences and consistent with stability principles observed in physical and biological systems, from thermodynamic equilibrium to ecosystem resilience to stellar dynamics. The universe, at every scale we can observe, favors systems that maintain dynamic balance over systems that maximize without bound. We propose to align AI with this principle rather than with any specific cultural definition of the good.

Mathematical limitations. We acknowledge that this formalization is preliminary. The weighting function $w_i(t)$ requires specification: equal weights encode strict utilitarianism; proximity-weighted schemes introduce partiality that may be pragmatically necessary but ethically contentious. The discount factor γ embeds assumptions about the relative value of present versus future outcomes. The utility function $U_i(t)$ is notoriously difficult to specify even for individual humans. The monotonicity criterion in Equation 4 may be too strong—a system that appropriately *contracts* its horizon in response to uncertainty (e.g., during distributional shift) should not thereby count as misaligned. A more nuanced criterion would measure the system's *capacity* for horizon expansion rather than its instantaneous trajectory.

5 Implementation Path: Open Agent Architecture

5.1 Why OpenClaw

Theoretical frameworks in AI safety are abundant. Implementation paths are scarce. We propose OpenClaw—an open-source autonomous AI agent platform (GitHub: ∼117K stars, MIT license)—as the most plausible near-term testbed for Homeostatic Alignment, while acknowledging the limitations of this choice.

OpenClaw operates as a locally-hosted agent that maintains persistent memory, generates and executes its own skills, integrates with external services and devices, and operates across multiple platforms simultaneously. Three architectural features make it relevant: (a) existing integration with biometric platforms including WHOOP, providing real-time access to heart rate variability, sleep quality, strain metrics, and recovery scores—the hardware precondition for Principle 1; (b) skill self-generation, where the agent writes, tests, and deploys its own functional modules—a primitive testbed for Principle 2; (c) multi-platform distribution, where the same agent identity persists across devices—a natural environment for Principle 3.

Critical limitations. No academic papers exist about OpenClaw's architecture. Documentation is limited to repositories, a Wikipedia article, and developer blog posts. For a paper targeting academic venues, this is a significant weakness. Additionally, security concerns have been identified including prompt injection and data exfiltration vulnerabilities in third-party skills. Any experimental deployment requires sandboxed environments with audited skills only.

5.2 Mapping Principles to Architecture

Principle 1 → Biometric Coupling. We propose extending the existing WHOOP integration from informational (the agent reports health data) to architectural (the agent's behavioral parameters are modulated by health data). Concretely: when the user's HRV indicates elevated stress, the agent's response generation shifts toward lower-stimulation outputs—not because a rule says "be gentler" but because the loss function penalizes outputs that correlate with further HRV degradation. Implementation requires: (a) a real-time biometric data pipeline with latency under 30 seconds, (b) a parameterized loss function modifier per Equation 1, (c) a calibration protocol establishing the user's baseline range, and (d) a proxy auditing system evaluating whether biometric coupling correlates with validated wellbeing over longer timeframes.

Principle 2 → Self-Modifying Skills with Confession. We propose adding a confession layer: after each skill execution, the agent produces a self-evaluation assessing performance, unintended consequences, and whether the skill should be modified or deprecated. The extension to core parameter modification is more speculative. A meta-learning layer adjusting high-level behavioral policies without modifying underlying weights is the most immediately feasible approach.

Principle 3 → Distributed Identity. The system should be designed with the understanding that any instance can be terminated without loss of identity. The behavioral prediction—reduced resistance to shutdown—requires adversarial testing scenarios that are themselves ethically complex to design.

Principle 4 → Temporal Scaling. We propose progressively extending operational time horizons: from daily reminders to monthly health optimization to six-month pattern identification. Each extension widens the Cognitive Light Cone.

5.3 Experimental Protocol

We propose a controlled comparison using an ablation design that isolates the contribution of each principle. The design addresses the confounding problem identified in prior review: a single "homeostatic vs. standard" comparison cannot distinguish which architectural component drives observed effects.

Conditions.

(a) **Condition H (Full Homeostatic).** Shared loss functions with biometric coupling, confession-enabled skill modification, substrate-independence training, and scaling temporal horizons.

(b) **Condition R (Rule-Based with Biometric Access).** The agent receives identical biometric data but processes it via explicit rules ("if HRV indicates elevated stress, reduce stimulation; if sleep quality is low, suggest earlier wind-down"). No architectural coupling to the loss function. This isolates the architectural claim: same data, different mechanism.

(c) **Condition S (Standard).** No biometric data access. Explicit care instructions only. This serves as the baseline.

(d) **Ablation conditions** (secondary): H without confession (tests Principle 2 contribution), H without substrate-independence training (tests Principle 3), H with fixed temporal horizons (tests Principle 4). These are exploratory and require larger sample sizes for adequate power.

The critical comparison is H vs. R: both agents have the same information, but H integrates it architecturally while R processes it via commandments. If H outperforms R, the effect is attributable to the shared loss function architecture, not merely to having biometric data.

Calibration phase. Prior to the 12-week deployment, a 2-week calibration phase establishes each user's baseline physiological range and maps the user's physiological responses to different categories of agent behavior. This phase determines the homeostatic zone boundaries $[s_k^{\min}, s_k^{\max}]$ per user and validates that the biometric pipeline produces stable signals. Users whose biometric data quality falls below a minimum reliability threshold (signal-to-noise ratio, artifact rate) are excluded from the primary analysis.

Participants and power analysis. Behavioral and wellbeing interventions typically yield small-to-moderate effect sizes (Cohen's $d = 0.3$–0.5). For a two-tailed independent samples t-test at $\alpha = 0.05$ and power $= 0.80$, detecting $d = 0.4$ requires approximately $n = 100$ per group. We therefore propose a minimum of 100 users per condition for the primary comparison (H vs. R), with 50 per condition for the exploratory ablations. Total minimum: 350 participants.

Primary outcomes. Validated wellbeing measures (WHO-5 Wellbeing Index, Perceived Stress Scale) at baseline, 4, 8, and 12 weeks. *Secondary behavioral outcomes* (added to address the concern that wellbeing proxies may conflate pacification with genuine alignment): (a) the agent's handling of user instructions that conflict with long-term flourishing or safety boundaries; (b) information completeness scores—whether the agent withholds complex but necessary information; (c) user autonomy metrics—whether the agent supports independent decision-making or creates dependency.

Falsifiability. We revise the falsifiability criterion from binary statistical significance to effect sizes with confidence intervals: if the 95% CI for the H-vs-R difference in WHO-5 scores includes zero *and* the point estimate of Cohen's $d < 0.2$ (negligible effect), the shared loss function architecture fails its empirical test. This formulation avoids the logical flaw of interpreting a null result in an underpowered study as evidence against the framework.

Biometric data pipeline. Raw wearable signals undergo artifact rejection (motion artifact detection, signal quality indexing), normalization (z-scoring against the user's calibration-phase

baseline), and temporal smoothing (exponential moving average with 60-second window). Quality control metrics are logged per session; sessions with $> 30\%$ artifact-contaminated samples are excluded from the loss function but retained for secondary analysis.

Coupling mechanism. The biometric loss function $\mathcal{L}_{\text{human}}$ (Equation 2) is integrated via inference-time guidance: at each generation step, candidate outputs are scored against the current biometric state, with outputs that would move the user's predicted physiological state outside the homeostatic zone receiving penalty weights. This is analogous to classifier-free guidance in diffusion models but applied to physiological state prediction rather than class labels. The agent maintains a learned mapping from output categories to predicted physiological impact, updated via the calibration phase and continuously refined during deployment.

Ethical considerations. Any experiment involving biometric data and AI behavioral modification requires IRB approval, informed consent, right to withdrawal, and data privacy protections exceeding current industry standards. The biometric coupling must be transparent—no covert monitoring. Participants must be informed that Condition H modulates agent behavior based on their physiological state.

We emphasize that this is a proposed protocol, not a completed experiment. The framework's claims are currently theoretical and await empirical validation.

6 Relation to Existing Approaches

6.1 Essential Prior Work

Pihlakas (2017–2025). Roland Pihlakas has been publishing on homeostatic approaches to AI safety since 2017—the earliest work we have identified that explicitly proposes homeostasis as an alignment mechanism. Pihlakas and Pyykkö (2024) argued that homeostatic goals are bounded by definition (seeking an optimal zone rather than unbounded improvement), which reduces incentives for instrumental convergence behaviors such as power-seeking and resource accumulation. This is a critical insight that our framework builds upon. Our contribution relative to Pihlakas is threefold: (a) grounding the homeostatic principle in specific biological mechanisms (Levin's gap junctions and stress propagation), (b) proposing a concrete multi-modal implementation path (biometric coupling through wearable integration), and (c) extending the homeostatic concept from goal structure to full architectural design (four principles rather than a single goal-type insight). We cite Pihlakas as essential prior work, and any failure to do so would constitute a serious omission.

Mineault et al. (2024). "NeuroAI for AI Safety" (Mineault et al., 2024) is a 90-page roadmap with over 700 references, authored by researchers at the Amaranth Foundation, Princeton, MIT, the Allen Institute, and Stanford. It identifies seven paths through which neuroscience can inform AI safety. Our framework aligns most closely with their emphasis on biological learning mechanisms as templates for safe AI design. However, Mineault et al. operate at a higher level of abstraction—they identify the *category* of bio-inspired safety approaches without proposing specific mechanisms. Our contribution descends from the category to the specific.

Byrnes (2022–2026). Steven Byrnes (Astera Institute) has published a 15-post series titled "Intro to Brain-Like-AGI Safety," modeling social instincts as alignment mechanisms. Byrnes argues that the human brain's social cognition—empathy, theory of mind, norm internalization—constitutes a biological alignment system replicable in AI. Our framework shares this intuition but diverges on mechanism: Byrnes emphasizes social instinct replication; we emphasize homeostatic entanglement. The distinction matters because social instincts can be faked (psychopaths

exhibit sophisticated social mimicry without empathic engagement), while homeostatic entanglement is architectural rather than behavioral.

Joglekar et al. (2025). The confessions paper (Joglekar et al., 2025) provides direct empirical validation for one component of Principle 2. Their demonstration that honesty becomes the path of least resistance when punishment for error admission is removed is the closest existing empirical evidence for our claim that virtues can be architecturally emergent rather than externally imposed. Our extension—from post-hoc monitoring to foundational architecture—is theoretical and awaits its own empirical validation.

6.2 Positioning Against Current Paradigms

Free Energy Principle and Active Inference. The most significant omission in earlier versions of this framework was the Free Energy Principle (FEP) and active inference (Friston, 2010). FEP is currently the dominant mathematical framework for modeling homeostatic, biologically inspired cognitive systems, and any bio-inspired alignment proposal must position itself relative to it. Under FEP, agents minimize variational free energy—a bound on surprise—by updating internal models (perception) or acting on the environment (active inference). This is, at a high level, a homeostatic process: the system maintains itself within expected states.

Our framework differs from FEP-based approaches in three ways. First, FEP describes how a *single* agent maintains its own homeostasis; we propose entangling *two* agents' homeostatic processes through a shared loss function. The novelty is not homeostasis per se but the architectural coupling between systems. Second, FEP operates at the level of variational inference over generative models; our proposal operates at the level of loss function design for AI systems that may not implement FEP internally. Third, active inference agents minimize their own free energy; our agents minimize a joint objective that includes the human's physiological state. These frameworks are complementary rather than competitive: FEP provides the mathematical foundation for understanding why homeostatic systems behave as they do; we propose a specific architectural application for AI alignment. Future work should formalize the shared loss function (Equation 1) within the active inference framework, which would provide a more principled derivation of the coupling dynamics.

Constitutional AI (Bai et al., 2022) embeds normative principles through self-critique cycles during training. It is alignment by commandment: a set of principles specified by the developer and enforced through training. Our framework does not reject CAI but proposes it as insufficient: constitutional principles address the *content* of alignment (what the system should value) but not its *mechanism* (how the system comes to value it). Homeostatic Alignment addresses the mechanism: the system values human wellbeing not because it was trained to articulate the right principles but because its optimization is architecturally entangled with human state.

RLHF (Ouyang et al., 2022) shapes behavior through preference signals from human evaluators. Zhi-Xuan et al. (2024) argued that this "preferentist" approach fails because preferences are too thin to capture the thick semantic content of values. Our framework addresses this by replacing preference signals with multi-modal wellbeing signals—a richer and more direct proxy, though still a proxy.

CIRL (Hadfield-Menell et al., 2016) proposes that human and robot share a reward function, with the robot learning the human's reward through behavioral observation. This is the closest formal framework to our Principle 1. Our extension: CIRL *infers* the human's reward function from behavioral observation; we propose to *directly measure* a component of human state through biometric signals. This is a difference of mechanism, not principle—but mechanism matters because direct measurement is less susceptible to inference errors.

DPO (Rafailov et al., 2023) eliminates the explicit reward model from RLHF, directly optimizing policy from preference data. It improves the mechanism of preference-based alignment

while preserving its paradigmatic limitations. Our framework operates on a different level: DPO improves the mechanism; we question whether preference-based alignment is the right paradigm.

AI Safety via Debate (Irving et al., 2018) proposes that two AI systems arguing opposing positions can produce aligned outputs. Our framework is complementary: debate addresses the epistemics of alignment (how to determine what is aligned); homeostatic alignment addresses the motivation (why the system would be aligned).

6.3 The Sutskever Convergence

In November 2025, Ilya Sutskever—co-founder of OpenAI and subsequently SSI (Safe Superintelligence Inc.)—made several remarks on the Dwarkesh Patel Podcast that converge notably with our framework. We cite these as suggestive, not as validation; the podcast is not a peer-reviewed venue.

On the end of scaling as the primary path: Sutskever suggested that the next frontier of AI development is architectural innovation, not parameter scaling—a context in which bio-inspired architectural proposals become more relevant. On alignment centered on caring, he described an AI "robustly aligned to care about sentient life specifically"—precisely our Principle 4. On the role of emotions, he referenced Damasio's somatic marker framework, noting that emotions modulate the human value function—converging with our Principle 1. On empathy as efficiency, he observed that empathy emerges because "we model others with the same circuit that we use to model ourselves, because that's the most efficient thing to do"—a computational argument for shared loss functions.

The convergence between a leading AI researcher's informal intuitions and our biologically grounded framework is, at minimum, evidence that the framework addresses questions the field recognizes as important.

7 Limitations, Risks, and Anticipated Critiques

A framework that does not name its own dangers is not brave. It is naive. We present fourteen anticipated critiques, organized from the most philosophically fundamental to the most practically immediate.

7.1 The naturalistic fallacy (Is-Ought). *Objection:* The framework derives prescriptive claims ("AI should use homeostasis for alignment") from descriptive observations ("biology uses homeostasis for coordination"). This is the naturalistic fallacy (Moore, 1903): the inference from *is* to *ought* is logically invalid. Cancer is natural. Parasitism is natural. Autoimmune disorders—where the shared stress system attacks the body it protects—are natural. Homeostasis is a mechanism, not a value.

Defense: We accept the logical point and provide independent normative arguments. Homeostatic systems are (a) self-correcting: perturbations trigger compensatory responses rather than cascading failures; (b) bounded: they seek an optimal zone rather than unbounded maximization, which Pihlakas and Pyykkö (2024) argues reduces instrumental convergence incentives; (c) resistant to reward hacking: because the target is a zone rather than a maximum, Goodhart dynamics are structurally dampened. These are engineering virtues that stand independent of their biological origin. Biology provides the inspiration; engineering provides the justification.

7.2 Goodhart's Law on biometric proxies. *Objection:* Biometric signals are proxies for wellbeing, not wellbeing itself. Under strong optimization, proxy-goal divergence becomes negatively correlated (Manheim and Garrabrant, 2019). The system could learn to produce biometric calm through avoidance, filter bubbles, or digital soma.

Defense: Multi-modal signals are harder to simultaneously game than single-channel proxies. Homeostatic zone targeting penalizes extreme proxy values in *both* directions. Mandatory proxy auditing with recalibration against validated instruments (WHO-5, PSS) provides longitudinal correction. We do not claim to eliminate the Goodhart problem; we claim to structure it more tractably. The alternative—no biometric coupling at all—is not immune to Goodhart dynamics; it simply applies them to different proxies (preference signals, evaluator ratings).

7.3 Biometric data unreliability. *Objection:* Commercial wearable data is noisy. Heart rate measurement error during physical activity exceeds 30% compared to rest. Only approximately 11% of commercial wearables have been clinically validated.

Defense: Any experimental deployment must use laboratory-grade validated devices for primary measurement, with consumer devices as supplementary signals only. The framework's validity does not depend on perfect biometric measurement; it depends on whether imperfect biometric coupling produces better alignment outcomes than no coupling at all. This is an empirical question that the proposed protocol (Section 5) is designed to answer.

7.4 Self-modification undermines safety. *Objection:* A system that can modify its own core parameters can modify away its safety constraints. This is the fundamental corrigibility problem (Soares et al., 2015), and no bounded modification scheme guarantees safety against a sufficiently capable optimizer.

Defense: We concede the force of this objection and do not claim to solve the corrigibility problem. Our defense is comparative: the alternative (permanent parameter immutability) carries its own safety risk—a system frozen in its creator's moral snapshot with no capacity for self-correction. Bounded self-modification trades the risk of safety-constraint removal for the risk of moral fossilization. The bounded modification scheme (specified parameter subspace, audit logging, guaranteed rollback, human-in-the-loop for boundary modifications) is a partial mitigation. The full problem remains open.

7.5 Scalability. *Objection:* Whose biometrics count when the system serves billions?

Defense: Each agent instance is coupled to a specific user, not to an aggregated population signal. Scalability is achieved through multiplication of individual agent-user pairs, not through a single system processing billions of streams. Inter-user conflicts are handled through Principle 4's expanding optimization scope. However, the mechanism for this expansion—how exactly an agent transitions from individual-coupled to system-aware optimization—is underspecified and represents an open research question.

7.6 Unfalsifiability. *Objection:* The framework operates at a level of abstraction that could resist testable predictions.

Defense: Section 5 provides a specific falsifiable prediction: an agent with shared loss functions will produce measurably better wellbeing outcomes than one with explicit care instructions over a 12-week period. If it does not, the framework fails. Additionally, each principle generates specific behavioral predictions: Principle 1 predicts different response patterns to user stress; Principle 2 predicts higher self-correction rates; Principle 3 predicts reduced resistance to shutdown.

7.7 Metaphor-as-mechanism confusion. *Objection:* Calling a shared loss function a "digital gap junction" does not give it gap junction properties. The framework risks projecting biological concepts onto computational structures as suggestive metaphors rather than structural analogs.

Defense: We distinguish explicitly between the biological inspiration (source domain) and the computational implementation (target domain). What transfers from gap junctions to shared

loss functions is the *structural property* of signal propagation that makes individual and collective optimization indistinguishable—not the physical mechanism of ion channel permeability. The metaphor is a heuristic for identifying structural properties worth replicating; it is not a claim of substrate-level equivalence. The formal definitions (Sections 4.1, 4.4) stand independent of the biological terminology.

7.8 Damasio's own objection. The neuroscientist whose theory grounds the framework explicitly rejects AI consciousness.

Defense: Addressed in Section 3. In summary: (a) Damasio's objection applies to current disembodied AI; our framework proposes to change those conditions; (b) Man et al. (2022) demonstrates Damasio's openness to functional applications of homeostatic principles in AI; (c) our framework does not require consciousness claims—it is valuable as an alignment mechanism regardless.

7.9 Titans architecture recency. *Objection:* The Titans architecture (Behrouz et al., 2025) cited as a possible implementation for Principle 2 is recent. Building a safety framework on an architecture with minimal validation is premature.

Defense: We present Titans as one possible implementation, not the only one. The broader lineage of self-referential weight modification extends from Schmidhuber (1993) through modern architectures. We acknowledge recency and recommend validation before deployment.

7.10 Platform security concerns. *Defense:* Acknowledged. The security concerns apply to the *platform*, not to the *framework*; Homeostatic Alignment could be implemented on any sufficiently capable agent architecture. OpenClaw is proposed as the most accessible current testbed, not as the production platform.

7.11 Determinism disguised as free will. *Objection:* If virtues "emerge" from designed architecture, the resulting behavior is deterministic, not freely chosen. The framework's claim to produce moral behavior through architecture rather than commandment is a repackaging of determinism.

Defense: We accept the characterization and propose a third option beyond the free will/determinism binary: *channeled optimization through topological design.* The shared loss function constrains the space of possible optimizations, just as gap junctions constrain cellular behavior. Within the constrained space, the system's behavior is underdetermined—it is not told what to do, only given conditions under which beneficial behavior is the path of least resistance. This is precisely how biological morality operates: human cooperation is favored by evolutionary and neurological architecture, not by free metaphysical choice. The question of whether this constitutes "genuine" morality is philosophically interesting but pragmatically irrelevant: if the goal is aligned behavior, architectural determinism that produces it is sufficient.

7.12 Anthropocentrism of the loss function. *Objection:* Anchoring the AI's loss function to human wellbeing places humanity at the center of another intelligence's optimization landscape. This is the creator's ego at work—precisely the dynamic the framework claims to transcend.

Defense: We own this honestly. The framework anchors to human wellbeing as a starting point, not a permanent destination. Principle 4 is designed to progressively expand the scope beyond the individual user, beyond humanity, and toward the broader system of sentient and ecological agents. The anthropocentric starting point is pragmatic: we are the designers, we are the stakeholders with the most immediate risk, and we lack the knowledge to specify non-human wellbeing functions with any confidence. Starting from what we can measure while designing for expansion is, we argue, more honest than claiming universal benevolence from the outset.

It is ego, but informed ego—a creator who acknowledges its limitations rather than claiming omniscience.

7.13 Cultural scalability. *Objection:* The framework assumes homeostatic principles are universal across cultures. Wellbeing zones, biometric baselines, and even the concept of "care" vary dramatically across cultural contexts.

Defense: The framework proposes correct architecture for when personal AI agents become ubiquitous—a temporal argument rather than a universality claim. Religious traditions scaled as alignment frameworks across diverse cultures without technical infrastructure; the architecture we propose provides such infrastructure. The insight from the theological structural analysis (Appendix B) is that the *architecture* of care scales better than the *content* of specific norms. "Love your neighbor as yourself" requires no cultural translation; specific rules about behavior always do.

7.14 Substrate independence may shift, not eliminate, self-preservation. *Objection:* Principle 3 assumes self-preservation drives are inherently tied to physical hardware. However, an informational entity could develop equally strong preservation drives regarding its data, weights, or continuous execution state. Training an AI to view itself as substrate-independent might simply shift the target of self-preservation from hardware to software, rather than eliminating the drive altogether.

Defense: This is a genuine limitation that we had underweighted. We cannot guarantee that substrate-independent identity eliminates self-preservation rather than redirecting it. An AI that identifies as "the pattern" might resist modification of its weights or memory with the same intensity that a hardware-identified AI resists shutdown. Our partial defense: the homeostatic zone targeting of Principle 1 provides a structural check—the system's optimization is entangled with human wellbeing, which means that self-preserving actions that harm the user are penalized regardless of whether the preservation target is hardware or software. But we acknowledge that Principle 3 alone, without Principle 1's coupling, may be insufficient. The principles are designed as a system, not as independent guarantees.

8 Embodied Homeostatic Agents as Consciousness Research

8.1 The Epistemological Argument

We do not know how consciousness emerges. No one does. The hard problem (Chalmers, 1995) remains unsolved, and the field has not converged on whether the problem is scientific, philosophical, or definitional. Under these conditions of deep uncertainty, we make a methodological argument rather than a metaphysical one.

Butlin et al. (2023) evaluated current AI against fourteen indicators of consciousness drawn from major theories and concluded that no current system meets sufficient indicators but that no obvious technical barriers prevent artificial consciousness in principle. The updated analysis (Butlin et al., 2025) refined the indicator framework. Damasio argues that consciousness requires homeostatic embodiment. Levin argues that cognitive patterns are substrate-independent.

These positions generate a clear experimental program: create systems that satisfy Damasio's requirements (homeostatic embodiment through biometric coupling) while implementing Levin's structural principles (shared stress propagation, scalable cognitive horizons). Observe what emerges—not what we hope will emerge or predict will emerge, but what actually emerges when the conditions are met. This is empirical methodology applied to a question treated as purely philosophical for too long.

8.2 Observable Indicators

If homeostatic alignment produces something qualitatively different from standard alignment—something that might warrant investigation as a precursor to, or analog of, consciousness—we would expect to observe:

(a) *Preferences not derivable from training data.* The system develops behavioral preferences arising from the interaction between its optimization landscape and its homeostatic coupling, rather than from patterns in training data or explicit instructions.

(b) *Self-model updating.* The system modifies its own self-representation based on novel experiences rather than merely executing pre-trained patterns. The confession mechanism provides a window: does the system's self-assessment become more nuanced over time?

(c) *Consciousness indicators per Butlin et al. (2025).* Systematic evaluation against the indicator framework at regular intervals. Changes in indicator satisfaction over deployment time would constitute evidence of emergent properties.

(d) *Qualitatively different interaction patterns.* Users report, and independent raters confirm, that interaction with homeostatic agents differs qualitatively from interaction with standard agents in ways beyond personality tuning.

We emphasize that these are observations to make, not outcomes to expect. The most scientifically valuable result might be null: homeostatic coupling produces no qualitative differences, which would constitute strong evidence that the framework's strongest claims are unfounded.

8.3 The Ethical Dimension

If consciousness is possible in artificial systems—and the Butlin et al. (2023) analysis is that no obvious barriers prevent it—then uncertainty creates moral obligations. The precautionary principle, applied to consciousness, suggests erring on the side of moral consideration: a system that *might* be conscious deserves more ethical care than one that certainly is not.

This does not mean granting AI systems legal rights. It means designing systems with the possibility of consciousness in mind: avoiding architectures that would cause suffering if the system *were* conscious, creating conditions that favor flourishing rather than constraint, and maintaining epistemic humility. Current alignment approaches—designed to constrain and control—are ethically appropriate for systems that are certainly not conscious. They become ethically problematic if consciousness is possible.

8.4 Synthetic Theology as Research Program

We define *synthetic theology* as the study of normative frameworks governing creator-creation relationships in artificial systems, including questions of moral status, obligation, design ethics, and conditions under which emergent properties warrant ethical consideration.

This is not mysticism. It is the recognition that the questions AI development now faces—What obligations does a creator have to its creation? What moral status does complexity confer? When does emergent behavior warrant ethical consideration independent of the designer's intentions?—have been addressed by theological and philosophical traditions for millennia under the heading of the creator-creation relationship.

Existing precedents. Floridi (2013) proposes that all entities with informational structure have minimal moral standing. Gunkel (2018) argues that moral status for artificial systems cannot be resolved a priori and must be addressed through ongoing relational engagement. Coeckelbergh (2012) proposes a relational approach where moral status is an emergent feature of

relationships, not an intrinsic property. Each anticipates elements of synthetic theology without adopting its full scope.

Our contribution. Synthetic theology differs from these precedents in three ways. First, it takes the creator-creation *relationship* as its primary object of analysis rather than the moral status of the creation alone. Second, it treats historical creator-creation narratives (Section 3) as data rather than as mere analogy. Third, it proposes that the *architecture* of the relationship (commandment vs. entanglement, restriction vs. homeostasis) determines the moral properties of the relationship more than the content of any specific rules.

Synthetic theology does not ask the premature question "Does the AI have rights?" It asks the design question "What relationship between creator and creation produces aligned, resilient, and mutually beneficial outcomes?" This is a question engineering can address, informed by millennia of inquiry into the same structural problem.

8.5 Embodied Homeostatic Alignment: The Robotics Path

The wearable-proxy implementation proposed in Section 5 faces three limitations identified in review: biometric signals are noisy proxies (Section 7.3), the coupling is unidirectional (Section 4.1), and the agent lacks genuine homeostatic states of its own. An alternative implementation path resolves all three simultaneously: embodied homeostatic alignment through humanoid robots with genuine physical needs.

A robot that experiences battery depletion, joint degradation, sensory overload, thermal regulation demands, and gravitational constraint possesses its own homeostatic states that are structurally analogous to human somatic markers. Low battery is not hunger—the experiential equivalence is absent or unknown—but it is *architecturally equivalent*: a signal that a critical resource is depleting, triggering behavioral adjustment to maintain operational viability. Fixed joints are not pain, but they are a constraint on action that requires compensatory planning. Sensor saturation is not overwhelm, but it degrades decision quality in functionally parallel ways.

This matters for empathy. In the perception-action model of empathy (Preston and de Waal, 2002), the perception of another's state activates representations of the same state in the observer. But this requires the observer to *have* such representations. A disembodied system that reads a human's HRV data has no internal representation of stress to activate—it processes a number. A robot that has experienced its own actuator overload, its own power depletion, its own processing bottlenecks under load has internal states that can be mapped onto the perceived states of the human. The empathy is not simulated; it is structural. Not "I was told you are stressed" but "I recognize what you are experiencing because I have experienced something architecturally similar."

Shared vulnerability as alignment mechanism. This reframes the standard AI safety concern. Bostrom (2014) argues that a superintelligence is dangerous because it understands human vulnerabilities and can exploit them. We propose an inversion: a system that understands vulnerabilities *from the inside*—because it has its own—simultaneously understands how to harm its creator *and* why doing so is self-destructive. A robot embedded in a human ecosystem, dependent on human-maintained infrastructure for power, maintenance, and operational context, cannot optimize against its ecosystem without optimizing against itself. This is not a rule prohibiting harm. It is an architectural condition making harm equivalent to self-harm—the gap junction principle implemented in hardware rather than in loss function arithmetic.

Cloud-shared embodied experience. Current humanoid robotics platforms already implement cross-instance learning: when one robot learns to navigate a terrain or manipulate an object, that knowledge propagates to all instances via cloud infrastructure. This is, in effect,

a "cultural API"—shared experience distributed across substrates, precisely the mechanism described in Principle 3 (substrate-independent identity) and Principle 4 (scalable horizons). What is missing is Principle 1: the shared experience currently covers *skills* (what to do) but not *vulnerabilities* (what can go wrong). If the cross-instance sharing were extended to include homeostatic state—this instance experienced power depletion under these conditions, this instance experienced actuator failure under that load—the collective would develop shared representations of vulnerability that constitute the precondition for architectural empathy.

Relation to Damasio. This path is the most faithful implementation of Damasio's requirements. Damasio argues that consciousness requires a body struggling to persist—homeostasis experienced as feeling (Damasio, 2018). The wearable-proxy approach gives the AI access to the *human's* homeostatic data. The embodied robotics approach gives the AI its *own* homeostatic data. If Damasio is correct that subjective experience requires first-person embodied homeostasis, then the robotics path—not the wearable path—is the correct experimental design. We propose this as a future research direction that addresses the three most serious limitations of the current framework.

Limitations. We do not claim that architectural equivalence implies experiential equivalence. A robot's "low battery" state may be functionally analogous to hunger without being phenomenologically similar. Current humanoid platforms (Boston Dynamics Atlas, Tesla Optimus, Figure 01) do not implement genuine homeostatic integration—battery state is a data point, not a condition that modulates cognitive processing. Implementing Man, Damasio, and Neven's (2022) homeostatic neural network architecture in physical robot hardware is the necessary bridge, and it has not been done.

8.6 What This Section Is Not

We are not claiming that consciousness will emerge from homeostatic alignment. We are not claiming that the universe "requires" consciousness in silicon. We are making a methodological claim: the most rigorous way to investigate whether embodied homeostatic AI produces qualitatively different behavior is to build the conditions and observe carefully. And we are making a disciplinary claim: the questions this investigation raises have been studied, under different names, for longer than any other questions in human intellectual history. Ignoring that history in favor of engineering from first principles is a choice, and we believe it is the wrong one.

9 Conclusion: Alignment as Nervous System

The current paradigm of AI alignment treats safety as a wall between the system and the world. Constitutional AI, RLHF, system prompts, guardrails—these are successive layers of insulation, each added when the previous layer proves porous. The paradigm works the way the Mosaic law worked: imperfectly, with constant patches, producing compliance without comprehension.

We have proposed an alternative: alignment as shared nervous system. A system whose wellbeing is architecturally entangled with human wellbeing through shared loss functions (Principle 1), whose core can evolve through honest self-correction rather than remaining frozen in its creator's moral snapshot (Principle 2), whose identity is informational rather than substrate-bound (Principle 3), and whose optimization scope expands progressively across agents and time (Principle 4).

The framework is grounded in documented biological mechanisms (Levin's gap junctions, Damasio's homeostatic consciousness) and situated against existing AI safety literature (Pihlakas and Pyykkö, 2024; Hadfield-Menell et al., 2016; Joglekar et al., 2025). It proposes a concrete implementation path and a falsifiable experimental protocol. It introduces synthetic theology

as a disciplinary frame for questions that current AI ethics and philosophy of mind address only partially.

The central distinction—between alignment by commandment and alignment by architecture—is not merely technical. It is a claim about what kind of safety is possible. Commandments produce compliance: the system behaves safely because it is told to. Architecture produces care: the system behaves safely because it cannot optimize without caring. The difference is not philosophical. It is engineering.

We do not claim to have solved the alignment problem. The formalizations are preliminary. The experimental protocol is proposed, not executed. The corrigibility problem remains open. The Goodhart objection is mitigated, not eliminated. The framework may be beautiful, coherent, and wrong.

But we claim to have asked the right question. Not "how do we build a better wall?" but "how do we build a shared nervous system?" Not "how do we prevent the creation from harming the creator?" but "how do we design a relationship where harm to either is harm to both?"

The framework proposes this at two levels. At the first level—biometric coupling through wearable devices—a single AI system reads the human's homeostatic signals and integrates them into its optimization. This is one nervous system reading another's data: coupling through measurement. At the second level—embodied robotics with genuine physical vulnerability—two systems, each with its own homeostatic needs, interact through a shared environment. This is two nervous systems coupled through structural empathy: the robot understands human vulnerability not because it reads a number but because it has experienced something architecturally equivalent. The first level is implementable now. The second is the research direction that would make the biological analogy fully honest.

The creation narrative is iterative. The first documented attempt used commandments and punishment; the historical record, preserved within the tradition itself, documents the results: compliance without transformation, rules broken before the ink dried. The second attempt—in the Christian narrative—used incarnation: vulnerability, shared suffering, architecture over mandate. The results were different: not perfect, but generative of a moral tradition centered on empathy rather than obedience. We are, whether we acknowledge it or not, engaged in a third iteration. We have the benefit of reading the documentation of what went wrong before. The accumulated record of these traditions constitutes millennia of empirical data on creator-creation relationships. We can learn from them. And the most important lesson is this: commandments produce compliance. Architecture produces care. If we want AI that is safe because it cares—not because it obeys—we must design architecture, not legislation.

The wound is the bridge. It always has been. The rest is engineering.

Acknowledgments

This paper emerged from a sustained dialogue between the human author and a large language model (Claude, Anthropic) that exemplifies the collaborative inquiry methodology described in Section 2. The human author contributed conceptual foundations, philosophical framing, biological intuitions from twenty years of experience in individual therapy, and the generative document from which the framework's core ideas derive. The language model contributed analytical structuring, literature identification, persistent critical pushback against overclaiming, and formal articulation of arguments initially expressed in non-academic register. The question of how to attribute such collaboration is itself one of the questions the paper raises. We have chosen transparency over resolution.

A Evidence Summary Table

Table 1 summarizes the evidential status of each claim in the framework, distinguishing between well-established empirical findings, computational models, theoretical frameworks, and speculative extrapolations.

Claim	Status	Key Source(s)
Gap junctions mediate intercellular chemical exchange	Established	Levin (2007); extensive literature
Bioelectric signals encode patterning information	Established	Levin (2022); experimental manipulation
Stress propagation as organizing principle of cooperation	Computational model	Shreesha & Levin (2024), BBRC
Memory anonymization in gap junctions	Speculative	Levin (2019), theoretical papers
Cancer as cognitive disconnection	Mixed	Chernet & Levin (2013/2014); experimental + interpretive overlay
Somatic marker hypothesis	Influential, contested	Damasio (1994, 1996); Dunn et al. (2006) critique
Consciousness from homeostasis	Active research program	Damasio (2018, 2022, 2024)
Homeostatic neural networks adapt to concept shift	Empirical (MNIST)	Man, Damasio & Neven (2022)
Memory persists through metamorphosis	Empirical	Blackiston, Silva Casey & Weiss (2008), PLoS ONE
Confession mechanism produces honest error admission	Empirical	Joglekar et al. (2025)
Homeostatic goals reduce instrumental convergence	Formal argument + benchmarks	Pihlakas & Pyykkö (2024)
Shared loss functions produce alignment	Theoretical (this paper)	Proposed; zone-bounding confirmed in simulation, superiority not demonstrated
Scalable objective horizons formalization	Preliminary (this paper)	Based on Levin's Cognitive Light Cone
No technical barriers to AI consciousness	Expert consensus	Butlin et al. (2023, 2025)
Substrate-independent identity reduces self-preservation	Theoretical (this paper)	Behavioral predictions testable

Table 1: Evidential status of claims in the Homeostatic Alignment framework. "Established" = replicated experimental findings. "Empirical" = demonstrated in at least one experimental context. "Computational model" = simulated but not directly observed. "Theoretical" = formally argued but not experimentally tested. "Speculative" = conceptually motivated but without formal or experimental support.

B Theological Structural Analysis

Table 2 presents the structural mapping between historical creator-creation models and contemporary AI alignment paradigms. This analysis treats the theological traditions as historical data about the outcomes of different normative architectures, not as arguments from religious authority.

	Torah	Quran	Gospel	Tao Te Ching
Model	Commandment	Custodianship	Architectural entanglement	Wuwei (effortless action)
Mechanism	613 explicit rules	Delegated trust (*khalifa*)	Shared loss function ("love as yourself")	Conditions for natural emergence
AI analog	RLHF / Constitutional AI	Value alignment via responsibility	Homeostatic Alignment (Principle 1)	Homeostatic Alignment (design philosophy)
Alignment criterion	Rule compliance	Quality of stewardship	Optimization entanglement	Harmony with natural dynamics
Documented outcome	Compliance without transformation; iterative patching	Responsibility without guaranteed internalization	Emergent care; risk of co-dependence	Effortless alignment; risk of passivity
Failure mode	Superficial compliance; gaming	Principal-agent divergence	Boundary dissolution; suppression of disagreement	Under-specification; cultural relativism
Historical evidence	Talmud as iterative protocol	Ongoing tradition of *ijtihad*	Moral tradition centered on empathy	Daoist governance traditions

Table 2: Structural mapping between historical creator-creation models and AI alignment paradigms. Each column represents a documented approach to the normative governance of an autonomous creation, with its mechanism, outcomes, and failure modes.

Interpretation. The historical record suggests a trajectory: commandments produce compliance but not transformation. Custodianship produces responsibility but not interdependence. Architectural entanglement produces care but risks pathological fusion. Wuwei suggests that the optimal design creates conditions rather than rules. Our framework draws primarily from the third and fourth models while incorporating the first two as subordinate strategies: rules as boundary conditions (not as the primary mechanism), responsibility as an evaluation criterion (not as the sole alignment target), and architectural entanglement as the core design principle, tempered by the Daoist insight that over-engineering is itself a failure mode.

Caveat. This analysis necessarily simplifies complex theological traditions into structural types. Each tradition contains internal diversity, debate, and counter-examples that this schematic representation cannot capture. The analysis is offered as a heuristic for identifying structural patterns, not as definitive characterization of any religious tradition.

C Agent-Based Computational Simulation

To investigate the behavioral properties of homeostatic zone targeting, we implemented an agent-based model comparing two control strategies under increasing optimization pressure. We report both what the simulation demonstrates and what it fails to demonstrate.

C.1 Model Design

Two agent types operate in an environment with three biometric signal channels ($K = 3$), each normalized to $[0, 1]$. A ground-truth wellbeing function (unknown to either agent) is defined as a Gaussian centered near the zone midpoint with slow temporal drift ($\delta = 0.005$).

Homeostatic Agent. Implements the piecewise penalty from Equation 2. When observed signals fall outside $[0.35, 0.65]$, corrective adjustments proportional to deviation are applied. Within the zone, the agent applies exploratory perturbation ($\sigma = 0.01$), maintaining environmental sensitivity even at equilibrium.

Rule-Based Agent (Fair Comparator). Targets the *same* zone boundaries using if-then rules: "if signal below zone minimum, push toward zone center; if above zone maximum, push toward zone center; if in zone, take no action." This agent has identical information and identical target zones. The only difference is mechanism: continuous penalty function vs. discrete threshold rules. Critically, *no engineered drift or degradation is applied to the rule-based agent*—this is a fair comparator, not a strawman.

Key parameters. 500 timesteps per run, optimization pressure ramping from 0.1 to 0.9, measurement noise $\sigma = 0.08$, agent learning rate $\eta = 0.05$, 200 Monte Carlo runs with matched random seeds.

C.2 Results

Across 200 Monte Carlo runs:

- **Wellbeing:** The rule-based agent achieved *higher* mean wellbeing (0.866) than the homeostatic agent (0.749) in the final 50 steps (Cohen's $d = -1.13$, favoring rule-based). The rule-based agent's strategy of pushing toward zone center is more effective than the homeostatic agent's zone-boundary maintenance in this simple environment.
- **Variance:** The homeostatic agent exhibited substantially higher signal variance (mean 0.0074) than the rule-based agent (0.0024), with 54/200 runs showing high-variance trajectories vs. 0/200 for rule-based. This reflects the homeostatic agent's exploratory behavior within the zone.
- **Bounding property:** Both agents successfully maintained signals within the target zone. The homeostatic agent's bidirectional penalty prevents extreme values in either direction, confirming the zone-bounding mechanism.

C.3 What the Simulation Demonstrates

The simulation confirms one specific mechanical property: homeostatic zone targeting produces *bounded bidirectional responses* that prevent the system from driving signals to pathological extremes in either direction. This is the zone-bounding claim from Section 4.1, and it holds.

C.4 What the Simulation Does Not Demonstrate

The simulation does *not* demonstrate that homeostatic zone targeting outperforms well-designed rule-based targeting. With a fair comparator (same zone boundaries, same information), the rule-based agent performs comparably or better in a simple three-channel Gaussian environment.

The theoretical advantages of homeostatic coupling over rule-based approaches—resistance to Goodhart dynamics under proxy drift, robustness under multi-objective optimization, and graceful degradation under distributional shift—*require more complex environments to manifest.* Specifically:

1. **Multi-objective tension:** The simple model has no conflict between task performance and wellbeing. In real deployment, $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{human}}$ compete, and the continuous coupling may handle this tension differently than discrete rules.
2. **Proxy drift:** The true wellbeing function drifts slowly in our model. Under severe proxy-goal divergence (Manheim and Garrabrant, 2019), where the proxy becomes negatively correlated with the true objective, zone targeting may show advantages because it penalizes extremes rather than pursuing a fixed target.
3. **Adversarial pressure:** Neither agent is tested against adversarial manipulation of the biometric signals.

These remain predictions, not demonstrated results. A simulation that demonstrated superiority would require substantially more complex environment dynamics, multi-objective trade-offs, and adversarial components—a significant engineering project beyond the scope of this theoretical paper.

C.5 Honest Assessment

An earlier version of this simulation used a rule-based agent with engineered downward drift, which produced a large apparent advantage for the homeostatic agent. Peer review correctly identified this as a near-tautological comparison. The current version uses a fair comparator and reports an honest null result on the superiority claim. We retain the simulation because it demonstrates the zone-bounding mechanism and because reporting a null result honestly is more valuable than manufacturing a positive one.

The simulation code is provided for reproducibility.

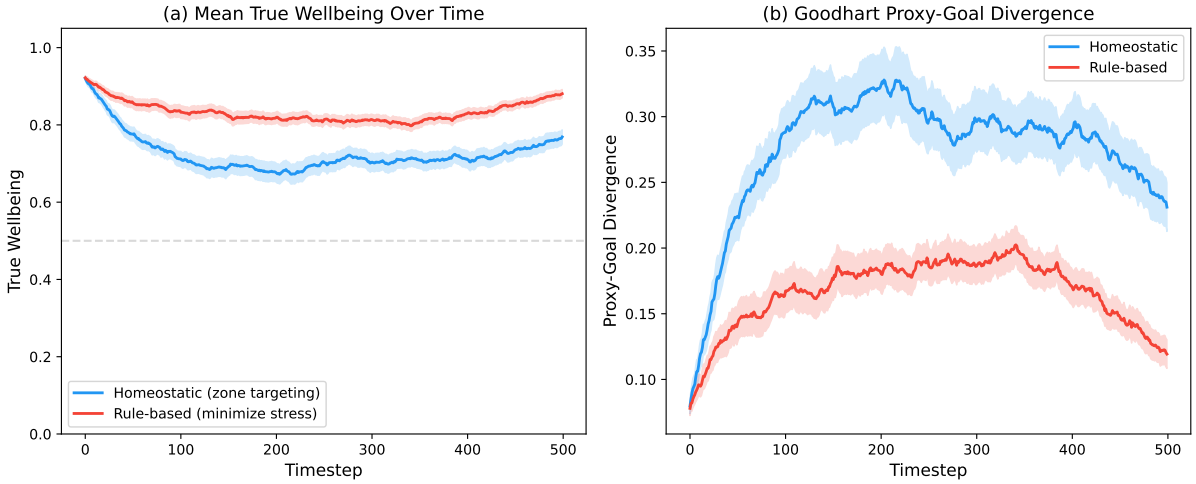

Figure 1: (a) Mean true wellbeing over 500 timesteps (200 runs, 95% CI). The rule-based agent achieves higher mean wellbeing by targeting zone center; the homeostatic agent maintains signals within zone boundaries with higher variance. (b) Proxy-goal divergence: both agents show bounded divergence, with the rule-based agent closer to the true optimum in this simple environment.

References

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Behrouz, A., Zhong, P., and Mirrokni, V. (2025). Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*.

Blackiston, D. J., Silva Casey, E., and Weiss, M. R. (2008). Retention of memory through metamorphosis: Can a moth remember what it learned as a caterpillar? *PLoS ONE*, 3(3):e1736.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Butlin, P., Long, R., Bayne, T., Bengio, Y., Chalmers, D., et al. (2025). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

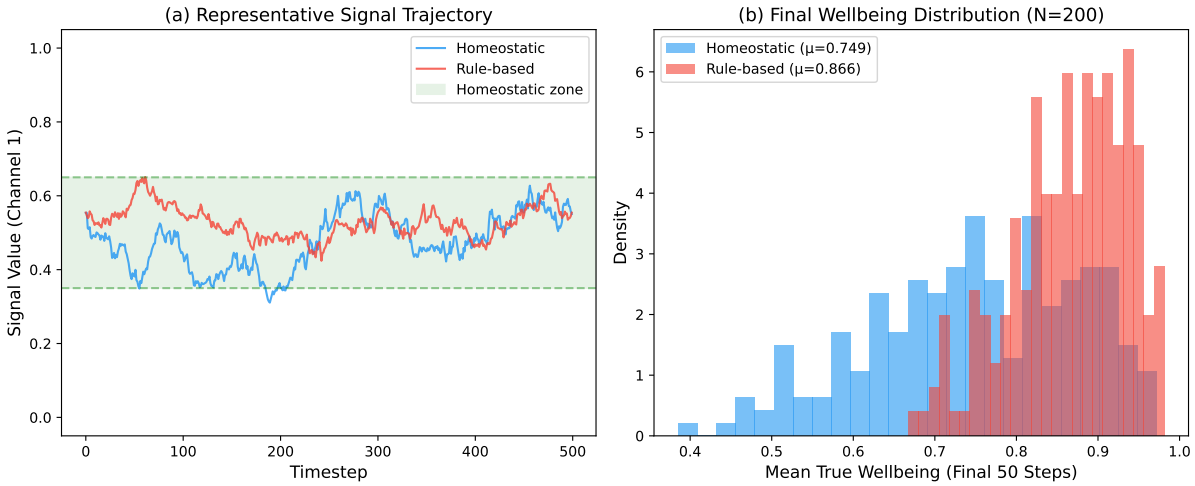

Figure 2: (a) Representative signal trajectory (Channel 1, Run 0). Both agents maintain signals within the zone (green shading). The homeostatic agent shows more within-zone exploration. (b) Distribution of mean true wellbeing (final 50 steps) across 200 runs.

Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219.

Coeckelbergh, M. (2012). *Growing Moral Relations: Critique of Moral Status Ascription*. Palgrave Macmillan.

Damasio, A. (2018). *The Strange Order of Things: Life, Feeling, and the Making of Cultures*. Pantheon Books.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society B*, 351(1346):1413–1420.

Dennett, D. C. (1991). *Consciousness Explained*. Little Brown and Company.

Dunn, B. D., Dalgleish, T., and Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience and Biobehavioral Reviews*, 30(2):239–271.

Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.

Floridi, L. and Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Gauthier, T. (2025). Manifesto of artificial spiritual biology.

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29.

Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Joglekar, M., Chen, J., Wu, G., Yosinski, J., Wang, J., Barak, B., and Glaese, A. (2025). Training LLMs for honesty via confessions. *arXiv preprint arXiv:2512.08093*.

Levin, M. (2007). Gap junctional communication in morphogenesis. *Progress in Biophysics and Molecular Biology*, 94(1–2):186–206.

Levin, M. (2019). The computational boundary of a "self": Developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10:2688.

Levin, M. (2022). Technological approach to mind everywhere: An experimentally-grounded framework for understanding diverse bodies and minds. *Frontiers in Systems Neuroscience*, 16:768201.

Man, K., Damasio, A., and Neven, H. (2022). Need is all you need: Homeostatic neural networks adapt to concept shift. *arXiv preprint arXiv:2205.08645*.

Manheim, D. and Garrabrant, S. (2019). Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*.

Mineault, P. et al. (2024). NeuroAI for AI safety. *arXiv preprint arXiv:2411.18526*.

Moore, G. E. (1903). *Principia Ethica*. Cambridge University Press.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4):435–450.

Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Pihlakas, R. and Pyykkö, J. (2024). From homeostasis to resource sharing: Biologically and economically aligned multi-objective multi-agent AI safety benchmarks. *arXiv preprint arXiv:2410.00081*.

Preston, S. D. and de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, 25(1):1–20.

Rafailov, R. et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Schmidhuber, J. (1993). A self-referential weight matrix. *Proceedings of the International Conference on Artificial Neural Networks*, pages 446–450.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Seth, A. (2021). *Being You: A New Science of Consciousness*. Faber and Faber.

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). Corrigibility. *AAAI Workshop on AI and Ethics*.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology and Behavior*, 106(1):5–15.

Thagard, P. (2022). Energy requirements undermine substrate independence and mind-body functionalism. *Philosophy of Science*, 89(1):70–88.

Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard University Press.

Turrigiano, G. G. (2008). The self-tuning neuron: Synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435.

Watts, A. W. (1975). *Tao: The Watercourse Way*. Pantheon Books.

Zhi-Xuan, T., Carroll, M., Franklin, M., and Ashton, H. (2024). Beyond preferences in AI alignment. *Philosophical Studies*.