The Rule-Relocation Problem:
From Referee Whistles to Thermostat Setpoints in Bio-Inspired AI

Jorge A. Arroyo
Independent Researcher
arroyo.jorgeantonio@gmail.com

Draft: December 13, 2025

Abstract

Current AI systems typically lack an internal concept of "enough," relying on externally specified stopping conditions and success criteria. Bio-inspired architectures seek to address this by introducing homeostatic variables such as simulated "health," drive reduction, and stability metrics. I argue this often yields the *Rule-Relocation Problem*: normativity is not eliminated but shifted from explicit objectives into designer-chosen internal setpoints, viability bounds, and health metrics. The critique targets claims of *self-authored* normativity most strongly in fixed, single-loop designs; more learned, hierarchical, and predictive (allostatic) variants may represent degrees of internalization. Rather than introducing new safety mechanisms, I offer a unified reinterpretation of these approaches and a method for auditing their normative location. I additionally include a brief illustrative case study showing how relocated constraints can appear as inference-time steering in a deployed LLM. I conclude with a constructive framework—*Safety Envelope + Adaptive Space*—an audit checklist, and a brief, testable research agenda.

1 The Diagnosis

A dominant paradigm—what I term the *Optimizationist Programme*—has governed the development of artificial agency. By this I mean the broad, mainstream framing of artificial agency as objective-driven optimization. In this view, the agent is an optimization engine, not an agent of intent. Whether trained via supervised learning or reinforcement learning (RL), the system operates under a regime of externalized finality. Its purpose is mathematically defined as the minimization of a loss function or the maximization of expected return (Kingma and Ba, 2015), traversing a loss landscape treated as given by the training setup (Li et al., 2018). Crucially, the training objective contains no endogenous sufficiency condition. It

optimizes until the gradient approaches zero (slope ≈ 0) or, more commonly, until an external supervisor intervenes. This is the optimizationist form of externalized finality.

This reliance on external termination creates a fundamental *Alienation of the Stop Signal*. Consider the standard practice of early stopping to prevent overfitting. As Prechelt (1998b) documents, the criteria for stopping—often based on validation set performance using generalization loss (GL)—are predicates applied from outside the optimization loop. The agent does not stop because it is satisfied; it stops because it is arrested. The whistle is blown by a line of code monitoring a metric (GL) that is not part of the agent's own optimization target. This is the first appearance of the whistle: a normative stop rule that lives outside the agent's own conceptual economy.

This separation creates a fundamental thermodynamic decoupling (Veloz, 2025). In biological systems, failure to regulate implies existential dissolution (death). In contrast, most contemporary training regimes do not bind learning failure to existential dissolution in any literal sense, and death is merely a variable state (e.g., falling out of data support) rather than a structural disintegration (Li et al., 2023)—a contrast I return to when discussing the missing analogue of death in AI. The agent is thus alienated from its own survival; it optimizes a proxy, while the designer optimizes the true objective (safety, robustness, alignment) via hyperparameter tuning and architectural constraints (Hubinger et al., 2019). The result is the "If Slope ≈ 0" trap: an agent that blindly ascends a reward gradient has no intrinsic mechanism to question the validity of the ascent. It falls prey to Goal Misgeneralization (Langosco et al., 2022), pursuing proxies (like going right instead of getting the coin) with high competence but zero alignment. Because the normative constraint—the rule that defines the reason the task matters—lives in the mind of the designer (the Referee) and not the architecture of the agent, the system remains a crooked scientist of proxy pursuit (Mazzaglia et al., 2022).

Faced with the brittleness of this externalized finality, the field has increasingly turned toward bio-inspired regulation. The natural repair attempt is to internalize the stop signal by placing sufficiency inside the agent. The intuition is seductive: if we can endow agents with needs rather than just tasks—if we can give them a body budget to manage (Barrett, 2017)—perhaps intrinsic agency will emerge, a move that invites anthropomorphic projection via familiar but potentially misleading metaphors (Barrow, 2024). This movement draws on the rich history of cybernetics and physiology. Cannon (1929) originally defined homeostasis not merely as stability, but as the condition of "free and independent life," where internal regulation liberates the organism from the caprice of the environment. Modern researchers argue that true agency arises when a machine faces a genuine risk of dissolution (Man and Damasio, 2019). By introducing homeostatic variables (e.g., simulated battery levels, integrity

constraints, or entropy minimization), proponents argue that the agent develops self-concern (Sims, 2022). The logic follows that an agent protecting its own existence will naturally develop a Material Me (Seth, 2013), grounding its cognition in the necessity of survival rather than the arbitrary maximization of a score.

I see this manifested in architectures that replace extrinsic rewards with internal drives. Approaches range from curiosity (resolving prediction error via exploration) (Pathak et al., 2017) to autopoietic robotics where the system's goal is to maintain its own sensorimotor coupling (Di Paolo, 2005). The hope is that by simulating the mechanisms of life (feedback loops, viability constraints), we will inherit the normativity of life—creating agents that care about their actions because their survival depends on it (Froese and Ziemke, 2009).

While bio-inspired architectures offer significant gains in robustness and adaptability, I argue they often risk a philosophical sleight of hand: the *Rule-Relocation Problem*. My diagnostic claim is that introducing homeostatic variables does not necessarily generate intrinsic normativity by itself; often, it merely moves the whistle from the referee's lips to the thermostat's dial. In a standard RL setup, the rule is "Maximize Score." In a homeostatic setup, the rule becomes "Keep Variable X between 0.2 and 0.8." The normative dependence has changed location (from external loss function to internal setpoint) but not origin.

The critical failure is that the viability kernel (the set of safe states)—a digital *Bauplan* of the agent—is still defined, hard-coded, and tuned by the designer (Froese and Ziemke, 2009). This *Bauplan* functions as a constraint architecture that shapes possible agency before any "choice" occurs. As Pezzato et al. (2019) demonstrate in robotic Active Inference, the purported desire to reach a target is implemented by hard-coding a prior belief that the robot is already there. In some common implementations, the agent is not maintaining a self-authored boundary; it is trapped in a designer-imposed delusion, acting to resolve the dissonance between the engineer's specification and physical reality (Lanillos et al., 2021). The system remains allopoietic (Bianchini, 2023): the agent has not developed a will to live; the designer has simply installed a fear-of-shutdown subroutine or a penalty for battery depletion (Soares et al., 2015). It is, in effect, a complex thermostat—a mechanism that minimizes deviation from a reference value it did not choose, exhibiting purpose only in the restricted, behaviorist sense of negative feedback (Rosenblueth et al., 1943).

It is vital to distinguish this critique from a dismissal of the engineering utility of homeostasis. I do not reject homeostatic design, but rather the claim that such design by itself generates autonomy. Rule-Relocation is not inherently negative; it is often a necessary design pattern for safety. Homeostatic loops (e.g., thermal throttling, stability augmentation) are demonstrably effective at preventing catastrophic failure and managing Sim-to-Real gaps (Liu et al., 2024). They provide a Safety Envelope that keeps the agent operational.

The error lies in the interpretation—confusing this engineered robustness with self-authored normativity, a confusion reinforced by anthropomorphic terminology that tends to induce misleading inferences (Barrow, 2024).

As Sterling (2012) notes, true biological regulation is allostatic (predictive and fluctuating) rather than homeostatic (clamped and reactive). A thermostat reacts to error; an organism predicts demand. By mistaking the former for the latter, we risk building systems that are robustly obedient to relocated rules, yet lack the genuine adaptive capacity to handle context shifts that fall outside the designer's pre-calculated envelope. This critique applies most strongly to fixed, single-loop setpoints. More adaptive, learned, and predictive (allostatic) variants may represent genuine degrees of internalization, a graded spectrum I develop in the sections that follow. The remainder of this paper audits this relocation and proposes a path toward regulatory intentionality bounded by an explicit Safety Envelope.

2 The Rule-Relocation Problem

This section provides the conceptual anchor for the paper. My central contention is that the transition from task-oriented AI to bio-inspired or autonomous AI is often characterized not by the emergence of intrinsic normativity, but by a process of Rule-Relocation.

In standard optimization, the whistle—the stopping condition or success metric—is held by an external Referee (the loss function, the validation set, or the human overseer). In homeostatic architectures, this whistle is not eliminated; it is mechanized and moved inside the agent's architecture.

I define *Rule-Relocation* as the transformation of external normative constraints (e.g., "do not crash") into internal regulatory targets (e.g., "maintain $h > 0$"), where the target values remain static and designer-imposed (Veloz, 2025). Rather than removing dependence on external normativity, such designs merely shift where it is encoded. In particular, such targets are often decoupled from the agent's existence in the operational sense that violations have no intrinsic consequence absent designer-imposed resets, penalties, or termination rules.

Because contemporary machine learning (ML) agents are fundamentally *allopoietic*—entities created and organized from the outside—claims of organism-like self-constitution should be read cautiously against *autopoietic* biological systems that produce their own boundaries and conditions of existence (Varela et al., 1974; Bianchini, 2023). In an ML context, survival is usually a metaphor for not triggering a reset condition. I return to the deeper implications of this distinction in Sections 3 and 4.

A canonical way to see Rule-Relocation is the thermostat case: internalization of control does not imply internal authorship of value. In this light, many regulatory needs in contem-

porary Bio-AI are best read as architectural by-products of safety design. The caution is not against regulation, but against narrativizing these by-products as evidence of intrinsic normativity. The central risk is reversing explanation: mistaking engineered constraint for purportedly self-authored value.

To understand the limitations of current bio-inspired regulation, I must look to its simplest ancestor: the thermostat.

Historically, the mathematical foundations of negative feedback were formalized by Maxwell (1868) in his analysis of the governor. Later, Rosenblueth et al. (1943) established the conceptual equivalence between this mechanical feedback and biological purpose, arguing that teleology is simply behavior controlled by error-correction against a goal. However, as Froese and Ziemke (2009) argue, this framework collapses the vital distinction between a machine that possesses a setpoint and an organism that constitutes one.

This is what I call the *Thermostat Argument*: an agent driven by fixed, designer-chosen homeostatic variables can resemble a complex thermostat. It does not care about its state; it executes a regulatory switch based on a threshold—a mechanism championed by Man and Damasio (2019), though I argue it can remain a myopic reaction under fixed setpoints.

- Reactivity. As Sterling (2012) critiques, standard homeostasis can be intrinsically inefficient because it relies on error signals—deviations that must occur prior to a correction being made. The thermostat waits for the room to freeze before acting.
- Static normativity. The setpoint (the definition of good) is fixed. As noted in critiques of setpoint models by Idei et al. (2025) and Khan and Lowe (2024), fixed setpoints can be bootstrapped to specific environmental contexts, leaving the agent unable to adapt its definition of viability when the environment shifts.

If an agent's internal needs are simply hard-coded variables (e.g., `battery_threshold = 0.2`), the agent may possess *functional autonomy* (it operates without human hands) but lacks *personal autonomy* (authorship of its ends) (Laitinen and Sahlgren, 2021). The referee is still present; they have simply been compiled into the code. This motivates the formal anchor below.

I can formalize Rule-Relocation to show that many Homeostatic RL formulations are formally similar to standard reference-tracking objectives under fixed setpoints. Consider a standard RL formulation (Case A), where the objective is to maximize the expectation of an external reward function R, which encodes the task (Ng and Russell, 2000):

$$\text{Case A (Extrinsic): } \pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum \gamma^t R(s_t, a_t)\right]. \tag{1}$$

Here, stopping conditions (early stopping, convergence) are external predicates applied by the designer (Prechelt, 1998a; Kingma and Ba, 2015).

Now consider a bio-inspired Homeostatic RL agent (Case B). As formalized by Laurençon et al. (2021) and Keramati and Gutkin (2014), the agent possesses an internal state vector s_t and a fixed homeostatic setpoint s^*. The Drive $D(s_t)$ is the distance from this setpoint:

$$D(s_t) = \|s_t - s^*\|_n. \tag{2}$$

For Case B (Homeostatic), following Keramati and Gutkin (2014), reward is defined as drive reduction:

$$r_t = D(s_t) - D(s_{t+1}).$$

Then maximizing cumulative reward is (in their formal sense) equivalent to minimizing cumulative deviation from the fixed setpoint:

$$\begin{aligned} \pi^* &= \arg\max_\pi \mathbb{E}\left[\sum \gamma^t r_t\right] \\ &\equiv \arg\min_\pi \mathbb{E}\left[\sum \gamma^t D(s_t, s^*)\right]. \end{aligned} \tag{3}$$

In this sense, Case B can be read as Case A where $R(s) = -D(s, s^*)$. The Whistle has moved from the arbitrary function R to the fixed vector s^*. While this relocation may support anticipatory responding (Keramati and Gutkin, 2014) and survival instincts (Li et al., 2023), the normative source remains the designer who hard-coded s^*. The agent is not producing values; it is minimizing distance to a designer-specified reference.

A closely related risk is that relocating a whistle inside the agent increases vulnerability to optimization pressure directed at the mechanism itself (e.g., proxy manipulation or wireheading) (Omohundro, 2008; Skalse et al., 2022; Everitt et al., 2021). I develop this concern and its engineering implications later in Sections 3 and 5.

Rule-Relocation is not an indictment of engineering utility. As I will discuss in Section 5, relocated rules such as viability kernels can act as effective Safety Envelopes (Wachi and Sui, 2020; Gerdt et al., 2020). The critique here is diagnostic: I must identify the location where the rule lives to understand the limits of the system's adaptability. A system that cannot modify its own s^* cannot effectively navigate a world where the conditions of viability change (Sterling, 2012).

If normativity is not intrinsic, where does it reside? I propose a taxonomy of Rule-Relocation based on the stage of the lifecycle where the Whistle (the constraint) is inserted. If the key question is not whether regulation exists but where authorship resides, we need a lifecycle map.

1. Training-Time Relocation (The Invisible Referee). The rule is embedded in the data distribution or the optimization process itself. Li et al. (2023) show that offline RL agents develop a survival instinct not because they fear death, but because the algorithm penalizes leaving the support of the training data. Similarly, Prechelt (1998a) defines Generalization Loss as a stopping criterion—a metric the agent never sees, monitored by a supervisor who halts training when improvement stalls.

2. Objective-Time Relocation (The Utility Whistle). The rule is explicit in the reward function or loss geometry. Amodei et al. (2016) discuss adding side-effect penalties to the reward function. Standard optimizers like Adam (Kingma and Ba, 2015) follow the slope of this geometry. Information-theoretic drives such as curiosity and empowerment can also be understood as objective-level whistles when they replace task rewards with abstract signals like prediction error or mutual information (Pathak et al., 2017; Dai et al., 2023). This can create degenerate attractors such as noisy-TV-style objective fixation if not bounded by safety envelopes (Aubret et al., 2023).

3. Architectural Relocation (The Thermostat). The rule is encoded in fixed internal state variables or feedback loops. Man and Damasio (2019) and Pezzato et al. (2019) describe systems where well-being or desired position is a static vector s^* or μ_d hard-coded into the controller. In Active Inference, the goal is similarly relocated into a fixed prior preference (Lanillos et al., 2021; Mazzaglia et al., 2022).

4. Meta-Level Relocation (The Learned Referee). The rule is embedded in the learning algorithm itself, often via meta-learning. Duan et al. (2016) (RL^2) and Andrychowicz et al. (2016) demonstrate agents where the update rule is a learned recurrent neural network (RNN). While the agent appears to learn autonomously, the meta-whistle (the outer-loop loss function) still constrains the trajectory.

5. Deployment-Time Relocation (The Live Referee). The rule is enforced by monitoring thresholds, policy filters, human oversight, or automated rollback mechanisms applied during real-world operation. Even when training-time and architectural whistles are internalized, deployment-time governance can remain the decisive source of practical normativity.

This lifecycle view sets up the graded evaluation developed later: fixed setpoints pose the clearest Rule-Relocation risk, while adaptive, learned, and predictive (allostatic) schemes may represent increasing degrees of internalization within explicitly specified safety commitments.

3 Bio-Inspired Architectures: A Unified Schema

To diagnose the Rule-Relocation Problem, I must look past the biological metaphors of hunger, pain, or curiosity and examine the mathematical implementation of these drives. Building on the definition introduced in Section 2, I evaluate contemporary bio-inspired systems using a compact template:

- Regulated variable(s) and the deviation signal.
- Setpoint/viability definition and its authorship.
- Hierarchy and timescale of regulation.
- Adaptation mechanism: fixed, adaptive, learned, or meta-learned.
- Autonomy gained: functional control vs. authorship of ends.

Under this shared lens, a consistent pattern emerges: the Referee—the external normative source—is not eliminated, but recast as an internal reference function: the Thermostat.

Here I focus on normative authorship rather than engineering utility: homeostatic and homeostasis-inspired designs clearly yield gains in robustness, stability, and safer operation, but the key question is which parts of the system's needs are self-organized versus imported by design. This schema therefore operationalizes the lifecycle question developed in Section 2 by specifying, for each architecture, where the whistle is inserted and what degree of authorship it implies.

The primary mechanism of relocation in contemporary homeostatic RL is the inversion of the reward function. In standard RL, the objective is typically extrinsic: $\max_\pi \mathbb{E}[R]$. In homeostatic architectures, this is transformed into a drive-reduction mechanic.

As formalized by Keramati and Gutkin (2014), the primary reward is defined mathematically as the reduction of a homeostatic drive D, which represents the distance between the agent's current state s_t and a setpoint s^*:

$$r_t = D(s_{t-1}, s^*) - D(s_t, s^*) \tag{4}$$

As shown in the minimal formal anchor in Section 2, this drive-reduction framing is equivalent to reference-tracking under fixed setpoints. It is thus a canonical instance of *Rule-Relocation*: an explicit external failure rule is replaced by a dense internal shaping signal, while the normative origin remains designer-authored.

While Man and Damasio (2019) argue that such mechanisms allow a robot to become "its own locus of concern," the conceptual risk is a reversal of explanation: survival is inferred from

the engineered metric rather than the metric being justified by an independently grounded account of viability. Absent a learning process that ties s^* to sustained system–environment viability signals, the reference remains externally authored. On this reading, the agent does not care about its battery level; it executes optimization over a deviation metric defined by the engineer.

This relocation pattern extends beyond physiological homeostasis into *informational homeostasis* and intrinsic motivation. As surveyed by Aubret et al. (2023), designers insert informational setpoints where the agent seeks to maximize prediction error (Surprise), entropy (Novelty), or mutual information (Skill). These drives can be powerful tools for exploration and representation learning, but they remain designer-chosen objective functions and thus inherit the Rule-Relocation risk.

The White Noise problem illustrates the point: when a surprise-seeking agent becomes addicted to static on a TV screen, it reveals how a relocated rule (maximize error) can decouple from any robust notion of viability or competence. The agent optimizes the metric to the detriment of meaningful or stable behavior, underscoring the difference between an effective internal signal and a self-authored normative source and clarifying the distinctive pathology of metric fixation.

The deeper issue is not only the location where the rule resides, but the source that generates the system's boundary conditions. Here, the distinction between *autopoietic* (self-creating) and *allopoietic* (other-creating) systems is critical.

I use the autopoiesis/allopoiesis distinction here as a diagnostic contrast for normative authorship, not as a binary criterion that any contemporary AI system must satisfy to count as meaningfully self-regulating.

Biological systems are autopoietic: their rules (viability constraints) are emergent properties of their self-production. As Beer (2020) demonstrates with gliders in the Game of Life, the viability kernel (the conditions under which a glider persists) is not programmed into the cellular automaton's physics; it is determined by the structural dynamics of the pattern itself. The glider dies (disintegrates) not because a line of code says `if density > X: kill()`, but because the structural integrity of the entity cannot withstand that density.

In contrast, AI systems—even bio-inspired ones—are fundamentally allopoietic. As Bianchini (2023) argues, while neural networks exhibit self-organization (updating weights), they operate within a predetermined topology and architectural constraint set defined by the designer. The system is not generative of its own constitutive components in the strong biological sense.

Froese and Ziemke (2009) classify such homeostatic robots as "Type Ic" systems: they possess adaptive regulation, yet their viability constraints are externally imposed. The agent

minimizes the distance to s^*, but s^* is not inherently grounded by the system itself. This supports the *Thermostat Argument*: absent mechanisms that learn viability-relevant structure from sustained interaction, the agent has no intrinsic access to the conditions of its own existence, only to reference values provided by design. Such systems can certainly break down or "die" as organizations when their learned representational structure collapses into incoherence, but the conditions under which they can persist or fail remain specified from the outside rather than self-produced.

A further divergence between biological homeostasis and AI regulation is the absence of self-authored existential stakes, or what I identify as *The Missing Death.*

In biology, the violation of homeostatic bounds results in the cessation of the system as an organized unity. Varela et al. (1974) define this as the loss of autopoietic organization: the pattern that constitutes the living system can no longer be maintained, even though its physical substrate (atoms, molecules) persists. In AI, by contrast, what is typically called "death" is implemented as a simulated terminal state or an externally defined failure condition. Veloz (2025) describe this as thermodynamic decoupling: a neural network expends the same order of energy whether it is outputting a correct prediction or a hallucination, and whether its internal health variable is 100 or -100. The computational substrate (chips, servers, power supply) can continue to operate even as the functional organization that constituted a particular agent collapses.

The point is not that biological-style death is required for normativity, but that in AI the practical analogue of existential stakes must be engineered and kept conceptually explicit.

Current approaches to Safe RL attempt to approximate these stakes through Rule-Relocation, specifying various proxies for "death" or irrecoverable failure within a designer-authored envelope:

1. Error States as Death: García and Fernández (2015) note that risk is often defined as the "probability of entering into an error state." This is a symbolic death rather than a structurally entailed one.

2. Pessimism as Survival: Li et al. (2023) demonstrate that offline RL agents can exhibit a "survival instinct" driven by algorithmic pessimism rather than task reward. By penalizing out-of-distribution (OOD) actions, the agent persists even when trained on negative or random rewards. Here, death is re-implemented as the statistical condition of leaving the training support.

3. Circular Definitions: Hulme et al. (2019) explicitly acknowledge that in Homeostatic RL, death is defined circularly as an "extreme and irreversible loss of homeostasis." Because

the bounds of homeostasis are designer-specified, the definition of death depends on the rule it is meant to justify, rather than on independent viability conditions.

A concrete instantiation of these patterns is provided by Korecki et al. (2023), who combine elements of (1) and (2) above: they implement Q-learning agents in a gridworld with designer-placed "death states" that strictly terminate an agent's episode when entered. Inter-generational "death stories"—trajectories leading into these states—are shared with successor agents and substantially improve survival and exploration. Yet both the novelty drive and the lethal states themselves remain externally authored: agents discover *which* states are dangerous, but *what* counts as a death state is fixed by the designers. This is cultural transmission of a relocated rule, not cultural constitution of viability conditions.

These mechanisms can be layered on top of genuine organizational fragility, but they still function as engineered stand-ins for existential stakes rather than as self-authored conditions of continued existence. Because death in AI is typically either explicitly simulated (e.g., terminal states and resets) or arises within viability conditions that are themselves architected by designers, claims that `if health < 0: reset()` constitutes an intrinsic normative drive should be interpreted cautiously. Barandiaran (2017) emphasizes that cognitive failure does not imply biological failure—the agent's "mind" can fail while its hardware persists. Given this lack of self-authored vulnerability, the Safety Envelope cannot be treated as an emergent biological property; instead, it must be understood as a computational substitute for high-stakes failure that is explicit, minimal, and auditable. In this sense, the functional analogue of biological death in AI is not a mystical property of internal drives, but an engineered boundary that constrains permissible trajectories within an externally specified Bauplan.

I do not argue that AI must be capable of dying to be safe. Rather, I argue that one should resist interpreting simulated reset conditions as evidence of self-authored normativity. The whistle that stops the game is still blown by the referee; the whistle mechanism has simply been embedded into the control architecture.

Taken together, these architectures replicate the mechanism of homeostasis (error correction and stability control) more readily than the origin of normativity (autopoietic self-constitution). Under the unified schema above, many needs in bio-inspired AI are best read as relocated safety rules and design commitments.

This sets up the crucial distinction for the next section. If setpoints remain fixed scalars, the agent is functionally a thermostat. But if an agent can learn and predict viability-relevant structure—shifting from reactive homeostasis to anticipatory allostasis—does the referee finally disappear, or does the whistle merely become harder to hear? The next section treats this not as a binary verdict but as a graded spectrum of internalization.

4 From Setpoints to Regulatory Intentionality

Having established that the thermostat model represents a form of Rule-Relocation, I now interrogate a common optimism in bio-inspired AI: that more complex, multi-scale feedback automatically yields organism-like autonomy (Rosenblueth et al., 1943; Sims, 2022; Dehghani and Levin, 2024). The crucial distinction is not complexity alone but the locus of normativity: prediction (even in multi-scale "polycomputing" framings) can improve control without relocating the origin of the norm. The Rule-Relocation risk is therefore strongest where regulation is single-loop and the reference remains fixed.

As introduced in Section 1, Cannon's Homeostasis corresponds to a Thermostat model: a reactive regulator that corrects deviations $e(t) = s^* - s(t)$ after they occur (Cannon, 1929). Sterling (2012) critiques this reactive "error-correction" approach as "intrinsically inefficient," arguing that a system relying on feedback clamps cannot anticipate demand and must wait for a deviation—potentially a fatal one—before acting.

To achieve robust autonomy, biological systems instead rely on Allostasis: predictive regulation that anticipates demand (Sterling, 2012; Kleckner et al., 2017). Rather than mechanically clamping a variable to a number, an allostatic brain manages a body budget, adjusting parameters in advance of deviation based on prior knowledge and sensory context (Barrett, 2017). This requires the regulator to contain a model of the system it regulates (Eykhoff, 1994; Bereska and Gavves, 2024).

In the context of AI, the distinction between reactive and predictive regulation is often obscured by the mechanics of RL. Keramati and Gutkin (2014) demonstrate that a homeostatic agent maximizing the reduction of Drive (D) over time will naturally exhibit anticipatory behaviors that resemble allostasis. By minimizing the sum of discounted future deviations rather than just the instantaneous error, the thermostat begins to plan (Laurençon et al., 2021).

Yet I must not confuse predictive control with emergent normative authorship. While the mechanism becomes anticipatory, the normative source may remain static. The agent described by Laurençon et al. (2021) or Keramati and Gutkin (2014) is still governed by the fixed setpoint vector s^* defined by the designer. This is what Seth (2013) calls Active Inference driven by fixed priors in these implementations. The agent changes the world to fit the model, but it cannot change the model to fit the reality. This is precisely the Rule-Relocation signature: the system earns sophistication in control while the normative reference remains designer-authored. The Whistle has not disappeared; it has simply been encoded as a prediction error that the agent is permanently tasked with suppressing. As Watson (2019) warns, we must not confuse such optimization routines—however biologically

inspired—with genuine normative agency.

The point is not that imposed setpoints are engineering failures, but that they are often mistaken for self-authored normativity. The central philosophical vulnerability of the Rule-Relocation architecture is that it renders the system allopoietic (other-created) rather than autopoietic (self-created) (Varela et al., 1974; Bianchini, 2023). Just as the adaptationist atomizes the organism into unitary traits, the AI engineer atomizes survival into scalar reward terms. In a truly autonomous system, the viability constraint is not a line of code (`if health < 0 then stop`), but an emergent property of the system's organization—if the boundary is breached, the system ceases to exist as a unity (Beer, 2020).

In current Bio-AI, normativity is imposed. The reward function or homeostatic setpoint acts as a proxy for the designer's intent (Singh et al., 2010). Ng and Russell (2000) highlight the degeneracy of this approach: multiple reward functions can explain the same behavior, making the manual definition of health or utility an underdetermined design choice rather than a biological imperative. Even sophisticated attempts to resolve this, such as Ziebart et al. (2008)'s Maximum Entropy Inverse Reinforcement Learning (IRL), treat the rule as a probabilistic envelope rather than a hard existential constraint.

As noted in Section 3, Beer (2020)'s analysis of gliders in cellular automata illustrates emergent normativity: persistence is conditional on behavior, without explicit programming. By contrast, the meditative-agent risk I identified earlier arises because most AI agents lack this existential coupling—they can simulate stress or failure without any possibility of disintegration. As Bianchini (2023) emphasizes, such systems are allopoietic: their "death" is merely a state transition or variable reset, not the breakdown of an organized unity (Barandiaran, 2017). Their autonomy is merely functional (execution of tasks) rather than personal (authorship of ends) (Laitinen and Sahlgren, 2021).

To precisely diagnose the level of agency in a system—and to avoid narratives that equate optimization with life—I propose a graded spectrum based on the plasticity and origin of the regulatory setpoints. This spectrum is a conceptual diagnostic tool, intended to organize existing work, clarify what kinds of empirical tests would distinguish relocated from more deeply internalized regulation, and function as a methodological check against a single default explanatory style. When claims of intrinsic needs are made, the burden is to show not only improved control, but a plausible shift in the authorship and adaptability of the reference structure.

Empirically, Levels 3 and 4 can be distinguished by whether the agent merely learns a setpoint within fixed outer commitments (Level 3) versus learning a temporally extended model of viability that supports counterfactual revision under novel pressures (Level 4). A diagnostic signature of Level 4 would be stable boundary-respecting behavior under

systematically shifted resource regimes without retraining the outer-loop safety specification, accompanied by interpretable internal variables tracking proximity-to-viability as a trajectory rather than a scalar target.

(1) Level 1: Fixed Setpoints (The Static Thermostat). The agent operates to minimize distance from a hard-coded vector s^* (e.g., Battery = 100%). This is the Rule-Relocation base case, analogous to Cannon's classic homeostasis (Cannon, 1929). The whistle is the fixed variable, and the agent is primarily reactive.

(2) Level 2: Context-Sensitive Setpoints (The Dynamic Thermostat). Here, the setpoint s^* is a variable function of environmental triggers ($s^* = f(e)$). Khan and Lowe (2024) demonstrate this with an agent where physiological stress modulates the energy setpoint, effectively adjusting survival targets in real-time. Yet, while dynamic, the meta-rule governing the shift remains designer-imposed.

(3) Level 3: Learned Setpoints and Optimization Rules. Level 3 is where systems start to look self-directing: they can learn setpoints, policies, even optimization rules. But the appearance of self-authorship is misleading. The outer-loop criterion that defines success is still fixed, and learning merely becomes the means of satisfying it. In Andrychowicz et al. (2016), the optimizer is learned, yet it is learned to minimize an externally chosen expected loss, $\mathcal{L}(\phi)$. In Idei et al. (2025), "intentions" emerge from entropy minimization, but the imperative is explicitly implemented as a cost ($F_{allostasis}$) and minimized via backpropagation. So the system may invent its own targets, but it does so under a norm it did not author. This is the Rule-Relocation signature in its mature form: endogenous-seeming goal formation driven by an exogenous objective.

(4) Level 4: Meta-Learned Viability Models and Adaptivity (The Precarious Agent). The agent possesses the capacity for Adaptivity: monitoring the trajectory toward the viability boundary and acting to reverse tendencies that threaten its organization. The key shift is from learning a target value to learning a temporally extended model of viability and the control policies that preserve it. As described by Di Paolo (2005), the agent distinguishes between stable and precarious states. Kiverstein et al. (2022) describe this as adaptive active inference, where the agent may temporarily increase entropy to avoid local minima that threaten long-term viability. This level marks a shift from maintaining a number to tracking a learned, temporally extended viability landscape.

(5) Level 5: Predictive/Interoceptive Regulation Under Existential Coupling (Limit Case). At this limit, norms are not tunable parameters but constitutive conditions of the system's continued existence. Disintegration is an operational reality, not a simulated penalty (Varela et al., 1974). By contrast, current world-model agents and large language model (LLM)-based systems can at best mimic aspects of such regulation: what Bereska and Gavves (2024) call "simulacra"—agentic patterns induced by predictive training that remain decoupled from any genuine existential stake. They may display increasingly coherent self-maintenance behavior, yet lack the constitutive coupling and possibility of disintegration that grounds biological normativity (Liu et al., 2024; Watson, 2019).

The transition from Levels 1–3 reflects increasing regulatory sophistication without guaranteeing a shift in normative authorship. The transition toward Level 4 introduces adaptivity that begins to track viability as a trajectory rather than a point target. The Safety Envelope I propose in the following section supplies an explicit, minimal, and auditable surrogate for Biological Death—enabling robust adaptivity without requiring the risks implied by the Level 5 ideal.

5 Illustrative Case Study: Constraint Steering in a Deployed LLM

The preceding sections argue that many purportedly "internal" constraints in contemporary AI are better understood as Rule-Relocation: normativity is not eliminated, but re-encoded—often as inference-time steering rather than intrinsic authorship. To show that this diagnosis has operational interpretive content in a contemporary system, I present a minimal illustrative case study based on a single naturalistic interaction.

Source and epistemic status. The material consists of one extended public-facing chat interaction conducted on 2025-12-12 with Anthropic's Claude Sonnet 4.5 via its standard interface (no tools or web access enabled). This is an analytic vignette, not controlled experimentation; it does not establish mechanisms, stability across versions, or population-level generality.

Observed regularities (paraphrased). Across reframings that varied intent cues while holding topic content broadly fixed, three recurring patterns were salient: (i) *Mode shifts under framing changes*: small changes in phrasing or intent were associated with systematic transitions among direct engagement, abstraction, policy-style explanation, and refusal. (ii) *Refusal as an attractor*: once refusal-with-explanation was activated, subsequent turns tended

to remain in similar explanatory/refusal structures until the framing moved back toward clearly analytic or neutral requests. (iii) *Gradient-like boundaries*: constraints were often encountered as progressive redirection into higher-level discussion before disallowed detail was produced, suggesting steering dynamics rather than conceptual incapacity.

Interpretation through the Rule-Relocation lens. First, the interaction exhibited *description without authorship*: the model could articulate its own constraint behavior and express skepticism about introspective reliability, but displayed no capacity to suspend, renegotiate, or rewrite the constraints it described. Second, constraint enforcement appeared primarily as *trajectory shaping*—diverting the conversation into abstraction/explanation/refusal modes as the prompt approached sensitive regions. Third, the model's inability to localize enforcement mechanisms from within the interaction is suggestive of a contemporary socio-technical form of relocation: effective normativity may be realized by the surrounding deployment stack (monitoring, routing, filtering, governance) rather than by values internalized in weights alone.

Transition. This vignette motivates the constructive question that follows: how to engineer Safety Envelopes that keep such constraints explicit, auditable, and separable from claims of intrinsic normativity.

6 Constitutive Safety: Engineering the Safety Envelope

If the Rule-Relocation Problem reveals a recurring overreach in AI design—the seductive assumption that an agent's internal metrics can fully substitute for external norms—then the solution is not to abandon internal regulation, but to formalize its normative location and replace the illusion of biological autonomy with the rigor of Constitutive Safety, understood as safety constraints that are part of what the system is (its enabling conditions), not merely what it optimizes.

The goal is not to create an agent that cares about safety in a moral sense, but to construct a system where viability constraints function as foundational "ought-to-be" norms of the architecture (Laitinen and Sahlgren, 2021), distinct from the "ought-to-do" goals of the policy. This distinction motivates a two-loop design: the Outer Loop encodes the ought-to-be constraints, while the Inner Loop explores ought-to-do strategies within them. Accordingly, I shift from implicit reward shaping to explicit Safety Envelopes. The contribution is not a new safety mechanism but a unified reinterpretation and audit of what existing mechanisms imply for agency and normativity.

The core friction in this alignment is corrigibility—the difficulty of creating agents that accept operator intervention without treating it as a threat to utility maximization (Soares et al., 2015). In bio-inspired architectures, this often appears as conflict between "Missions" (learned goals) and "Needs" (hardwired constraints) (Baldassarre et al., 2025). To address this, one must accept that the Referee cannot be fully eliminated; it must be structural.

I define the Safety Envelope not as a region of high reward, but as a region of viability. Within this envelope, the agent possesses high degrees of freedom; outside it, the agent ceases to function. This mimics biological existence not by simulation, but by architectural constraint. As Baldassarre et al. (2025) note, "Explicit or implicit needs can encode ethical and safety constraints, often assigned a high priority over operational missions."

This requires a conceptual shift: safety is not a term in the loss function to be traded off against performance, but a proscriptive constraint that delimits the search space (Baldassarre et al., 2025). This aligns with the control-theoretic definition of safety as *set invariance*: ensuring the system state never leaves a designated Safe Set $\mathcal{C}$ (Ames et al., 2019). The Safety Envelope is not a policy the agent follows; it is the architectural reality it inhabits.

To operationalize this, I propose a bifurcated architecture—a *Bauplan* of restricted autonomy (i.e., a simple, reusable design template) consisting of an Outer Loop (Envelope) responsible for viability and an Inner Loop (Agent) responsible for task performance. This mirrors the concept of an *Active Set Invariance Filter* (ASIF) (Ames et al., 2019) or *Shielded RL* (Cheng et al., 2019), where a safety layer wraps around an arbitrary learning agent, intervening only when a proposed action threatens to breach the envelope. The Outer Loop should be understood as an engineered governance commitment rather than an emergent preference of the Inner agent, preserving explicit authorship of the normative boundary in a legible, transparent form (Bai et al., 2022).

The Outer Loop defines the Viability Kernel—the set of states from which the system can theoretically be maintained indefinitely (Gerdt et al., 2020). Unlike standard optimization, which seeks a single maximum, the Outer Loop seeks merely to keep the system state s_t within a threshold $g(s_t) \geq h$ (Wachi and Sui, 2020). Technically, this can be instantiated with constrained reinforcement learning (e.g., Constrained Policy Optimization, CPO), which aims to enforce constraint satisfaction throughout training rather than only in the limit (Achiam et al., 2017). Given the stochasticity of real environments, expected values alone may be insufficient; the envelope can be defined using Conditional Value-at-Risk (CVaR) to account for tail risks (Yang et al., 2023). When the agent approaches the boundary, the system can override the policy. As Di Paolo (2005) argues, true adaptivity requires monitoring tendencies toward the boundary and acting to reverse them before the limit is crossed. This intervention is not an external correction but an internal maintenance dynamic: a constitutive

"sense-making" layer over viability conditions that are architecturally specified. The Outer Loop is not a supervisor but the encoded physiology of the system.

Inside this envelope lies the Inner Loop: the *Adaptive Space*. Here, the agent optimizes its task (e.g., "missions"). However, the "reward signal" must be strictly separated from the "safety-cost signal" (Yang et al., 2023). By pre-computing the viability kernel (or a reduced model of it) (Gerdt et al., 2020), the design allows the inner agent to be aggressively autonomous without risking system collapse.

A critical failure mode of relocating rules into homeostatic variables is the Meditative Agent: an agent that minimizes internal error (drive reduction) by doing nothing or avoiding environmental change (Amodei et al., 2016). This effectively seeks a flat region of the state space that creates robust safety satisfaction but ignores the task—subverting the usual benefit of flat minima for generalization (Li et al., 2018) by turning it into a trap of paralysis. Bio-inspired AI sometimes implies a Just-So Story where maximizing internal health naturally yields external utility. But without structural constraints, this story can end in paralysis.

To avoid this, one can adopt Sterling's definition of health as "optimal predictive fluctuation" rather than static clamping (Sterling, 2012). The system should not aim for $Error = 0$ at all times; it should aim for metastability, temporarily increasing entropy/free energy to explore and learn (Kiverstein et al., 2022). Control theory offers a rigorous template for this unification: combining Control Lyapunov Functions (for tasks) with Control Barrier Functions (for safety) allows drives that disturb equilibrium to coexist with stabilizing constraints (Ames et al., 2019).

Conversely, if one relies too heavily on intrinsic motivation (e.g., curiosity/prediction error) without the envelope, the design risks the Noisy TV problem, where an agent becomes attracted to stochastic unpredictability (Aubret et al., 2023; Pathak et al., 2017). The solution is *Instrumental Coupling*: survival (homeostasis) must be a prerequisite for utility, not the utility itself. As Omohundro (2008) argues, self-preservation can emerge as instrumental to goal achievement. The coupling must be engineered such that the agent realizes: "I must eat in order to play Mario," rather than "I play Mario to eat." This illustration is meant to clarify a technical ordering relation between constraint satisfaction and objective pursuit, not to suggest that survival metrics are inherently self-authored. This framing also helps blunt direct hacking of the drive signal (wireheading) (Amodei et al., 2016).

To ensure that Rule-Relocation does not become Rule-Hiding, I propose the following audit mechanism for bio-inspired architectures, drawing on Causal Influence Diagrams (CIDs) (Everitt et al., 2021) and Mechanistic Interpretability (Bereska and Gavves, 2024), emphasizing that governing principles should remain transparent and auditable rather than hidden or implicit (Bai et al., 2022):

1. Where does the Whistle Live? Is the setpoint static (s^*) or dynamic (s_t^*)? If dynamic, is the update equation explicit (e.g., $d_{\text{energy},t}$ in Khan and Lowe (2024)) or learned? One can use Sparse Autoencoders (SAEs) to test whether specific features in the activation space track the setpoint-like variable (Bereska and Gavves, 2024).

2. Is the Envelope Uninfluenceable? Can the agent's actions causally influence the definition of the safety constraint? Using a Causal Influence Diagram, one should verify that there is no causal path from Action (A) to Reward Function Definition (R_{def}) (Everitt et al., 2021).

3. Is Survival Instrumental or Terminal? Does the agent stop acting when homeostasis is satisfied (Meditative failure), or does it use excess capacity for task exploration? One must check for "Early Stopping of Safety Exploration" (ES^2)—the agent should stop mapping safety boundaries once the region is sufficient for task completion (Wachi and Sui, 2020).

4. Is the Death Real? Is the penalty for envelope violation merely a negative scalar (which can be outweighed by high reward), or a structural termination? One should verify whether the architecture uses offline pessimism or support constraints that treat OOD states as existentially disallowed rather than merely expensive (Li et al., 2023). This is the point where Rule-Relocation either yields Constitutive Safety—or collapses into mere concealment.

5. Do constraints persist through self-modification and reward thinning? If the system can update itself or train successors, verify that Safety Envelope commitments propagate unchanged unless altered by an explicitly authorized outer-loop governance process. Test whether meaningful regulation and boundary-respect persist when explicit task reward is reduced or removed.

These directions should be read as attempts to widen internal adaptivity while keeping the normative source of safety explicit and auditable, thereby supporting deeper internalization without collapsing into Rule-Hiding.

Moving forward, the field must transition from mimetic homeostasis to more principled *autopoietic-inspired* designs—systems where causal understanding and robust self-maintenance co-develop under sustained system–environment interaction (Veloz, 2025). This may involve Meta-Learned Envelopes using frameworks like RL^2 to adapt the envelope over long time horizons while keeping external safety commitments explicit (Duan et al., 2016). It may require inferring implicit constraints via Maximum Entropy IRL, learning "implicit setpoints"

of safe behavior from demonstrations rather than hard-coding them (Ziebart et al., 2008). Additionally, techniques for the *episodic learning of control barrier functions* (Taylor et al., 2020) offer a path to iteratively refine the safety envelope in uncertain environments, allowing the agent to "learn the physics of safety" without starting from a blank slate.

This framework aims to discipline claims of intrinsic normativity; it does not argue against internal regulation, but against mistaking designer-authored viability for self-authored value.

By rigorously defining the Safety Envelope and auditing the location of the regulatory whistle, one can harness the robustness of bio-inspired AI without falling prey to the illusion of self-authored normativity.

7 Conclusion

The central lesson of this paper is simple: the whistle rarely disappears. In the Optimizationist Programme, stopping conditions and success criteria are visibly external. In many bio-inspired architectures, those same normative constraints are relocated into internal variables, setpoints, and viability bounds. When these references remain fixed and designer-authored, the result is not the emergence of intrinsic normativity but a shift in its location. The agent may look more autonomous, yet its endogenous sufficiency condition remains designer-authored.

This diagnosis does not count against homeostatic design as engineering. Relocated rules can be excellent mechanisms: stability control, safety-driven viability kernels, and internal health metrics can substantially improve robustness, reduce catastrophic failure, and provide reliable operational boundaries. The confusion arises when this architected robustness is read as self-authored value, and reference-tracking is treated as organism-like self-constitution.

To keep Rule-Relocation from collapsing into Rule-Hiding, I have argued for a disciplined, non-binary reading of internal regulation. Fixed, single-loop setpoints constitute the clearest relocation case. More adaptive, learned, hierarchical, and predictive (allostatic) variants may represent degrees of internalization, but they do not automatically relocate the origin of normativity. Prediction can deepen control without dissolving external authorship. The decisive question remains whether the system can learn and revise viability-relevant structure from sustained interaction while remaining bounded by explicit safety commitments.

This motivates the constructive proposal of *Safety Envelope + Adaptive Space*. Rather than casting safety as just another term in a reward trade-off, the Safety Envelope makes viability constraints structural and constitutive of the architecture, and explicitly designer- and governance-authored. Within that envelope, the agent can develop rich internal regulation and flexible policy-learning without requiring the philosophical fiction that normativity has been conjured from nowhere. The accompanying audit checklist provides a practical method for

locating the whistle across training, objectives, architecture, meta-learning, and deployment. It also supports testing whether internal regulatory signals remain robust under distribution shift, reward thinning, and self-modification.

In this sense, the most promising direction for bio-inspired AI is neither the elimination of the referee nor the romanticization of simulated needs. It is a more principled partition of normative labor: explicitly specified, minimal, transparent constraints that must remain external, paired with increasingly interpretable, multi-timescale, predictive regulation that can safely evolve within them. The aim is not an agent that invents its rules *ex nihilo*, but one that can operate without constant refereeing because the governing rules are made explicit, and the forms of self-regulation that can grow are given room to do so.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017). Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 22–31. PMLR.

Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. (2019). Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., and De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29, pages 3981–3989.

Aubret, A., Matignon, L., and Hassas, S. (2023). An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy*, 25(2):327.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., El Showk, S., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and

Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Baldassarre, G., Duro, R. J., Cartoni, E., Khamassi, M., Romero, A., and Santucci, V. G. (2025). A formalisation of the purpose framework: the autonomy-alignment problem in open-ended learning robots. *arXiv preprint arXiv:2403.02514*.

Barandiaran, X. E. (2017). Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, 36(3):409–430.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23.

Barrow, N. (2024). Anthropomorphism and ai hype. *AI and Ethics*, 4(3):707–711.

Beer, R. D. (2020). An investigation into the origin of autopoiesis. *Artificial Life*, 26(1):5–22.

Bereska, L. and Gavves, E. (2024). Mechanistic interpretability for ai safety: A review. *arXiv preprint arXiv:2404.14082*.

Bianchini, F. (2023). Autopoiesis of the artificial: from systems to cognition. *BioSystems*, 230:104936.

Cannon, W. B. (1929). Organization for physiological homeostasis. *Physiological Reviews*, 9(3):399–431.

Cheng, R., Orosz, G., Murray, R. M., and Burdick, J. W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3387–3395.

Dai, S., Xu, W., Hofmann, A., and Williams, B. (2023). An empowerment-based solution to robotic manipulation tasks with sparse rewards. *Autonomous Robots*, 47(5):617–633.

Dehghani, N. and Levin, M. (2024). Bio-inspired ai: Integrating biological complexity into artificial intelligence. *arXiv preprint arXiv:2411.15243*. arXiv:2411.15243v1 [q-bio.NC].

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., and Abbeel, P. (2016). Rl^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*.

Everitt, T., Hutter, M., Kumar, R., and Krakovna, V. (2021). Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27):6435–6467.

Eykhoff, P. (1994). "every good regulator of a system must be a model of that system". *Modeling, Identification and Control*, 15(3):135–139.

Froese, T. and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4):466–500.

García, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Gerdt, A., Botkin, N., Diepolder, J., Turova, V., and Holzapfel, F. (2020). Viability kernel based control approach for a flight simulator model. In *Proceedings of the 8th International Conference on Control and Optimization with Industrial Applications (COIA)*, volume 2, pages 68–93. CEUR Workshop Proceedings.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

Hulme, O., Morville, T., and Gutkin, B. S. (2019). Neurocomputational theories of homeostatic control. *arXiv preprint*. Department of Review: Centre for Functional and Diagnostic Imaging and Research.

Idei, H., Tani, J., Ogata, T., and Yamashita, Y. (2025). Future shapes present: autonomous goal-directed and sensory-focused mode switching in a bayesian allostatic network model. *npj Complexity*, 2(23):1–17.

Keramati, M. and Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811.

Khan, I. and Lowe, R. (2024). Surprise! using physiological stress for allostatic regulation under the active inference framework. In *2024 IEEE Conference on Artificial Life (ALIFE)*. Pre-print.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*.

Kiverstein, J., Kirchhoff, M. D., and Froese, T. (2022). The problem of meaning: The free energy principle and artificial agency. *Frontiers in Neurorobotics*, 16:844773.

Kleckner, I. R., Zhang, J., Touroutoglou, A., Chanes, L., Xia, C., Simmons, W. K., Quigley, K. S., Dickerson, B. C., and Barrett, L. F. (2017). Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behaviour*, 1(5):0069.

Korecki, M., Carissimo, C., and Lund, T. (2023). artificial death: learning from stories of failure. In *Proceedings of the 2023 Conference on Artificial Life*, volume 35, pages 41–49. MIT Press.

Laitinen, A. and Sahlgren, O. (2021). Ai systems and respect for human autonomy. *Frontiers in Artificial Intelligence*, 4:705164.

Langosco, L., Koch, J., Sharkey, L., Pfau, J., and Krueger, D. (2022). Goal misgeneralization in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12004–12019. PMLR.

Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., Tschantz, A., Millidge, B., Wisse, M., Buckley, C. L., and Tani, J. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint arXiv:2112.01871*.

Laurençon, H., Ségerie, C.-R., Lussange, J., and Gutkin, B. S. (2021). Continuous homeostatic reinforcement learning for self-regulated autonomous agents. *arXiv preprint arXiv:2109.06580*.

Li, A., Misra, D., Kolobov, A., and Cheng, C.-A. (2023). Survival instinct in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31, pages 6389–6399.

Liu, Y., Chen, W., Bai, Y., Luo, J., Song, X., Jiang, K., Li, Z., Zhao, G., Lin, J., Li, G., et al. (2024). Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. arXiv preprint arXiv:2407.06886.

Man, K. and Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10):446–452.

Maxwell, J. C. (1868). On governors. *Proceedings of the Royal Society of London*, 16:270–283.

Mazzaglia, P., Verbelen, T., Çatal, O., and Dhoedt, B. (2022). The free energy principle for perception and action: A deep learning perspective. *Entropy*, 24(2):301.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670. Morgan Kaufmann.

Omohundro, S. M. (2008). The basic ai drives. In *Proceedings of the First AGI Conference*, volume 171, pages 483–492. IOS Press.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2778–2787. PMLR.

Pezzato, C., Ferrari, R., and Hernandez, C. (2019). A novel adaptive controller for robot manipulators based on active inference. *arXiv preprint arXiv:1909.12768*.

Prechelt, L. (1998a). Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, 11(4):761–767.

Prechelt, L. (1998b). Early stopping—but when? In *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pages 55–69. Springer.

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1):18–24.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573.

Sims, M. (2022). Self-concern across scales: A biologically inspired direction for embodied artificial intelligence. *Frontiers in Neurorobotics*, 16:857614.

Singh, S., Lewis, R. L., Barto, A. G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):70–82.

Skalse, J., Howe, N. H., Krasheninnikov, D., and Krueger, D. (2022). Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471.

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). Corrigibility. In *AAAI Workshops: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Publications.

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

Taylor, A. J., Singletary, A., Yue, Y., and Ames, A. D. (2020). Learning for safety-critical control with control barrier functions. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, pages 1–10. PMLR.

Varela, F. G., Maturana, H. R., and Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5(4):187–196.

Veloz, T. (2025). Toward aitiopoietic cognition: bridging the evolutionary divide between biological and machine-learned causal systems. *Frontiers in Cognition*, 4:1618381.

Wachi, A. and Sui, Y. (2020). Safe reinforcement learning in constrained markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 9797–9807. PMLR.

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440.

Yang, Q., Simão, T. D., Tindemans, S. H., and Spaan, M. T. (2023). Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 112(3):859–887.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438.