The Rule-Relocation Problem in Bio-Inspired AI

Jorge A. Arroyo
Independent Researcher

February 5, 2026

Abstract

Bio-inspired AI is increasingly presented as overcoming externally imposed objectives by internalizing regulation through drives, homeostasis, or predictive control. This Perspective argues that this narrative is often mistaken: in many contemporary designs, normativity is not eliminated but relocated, with constraints compiled into internal variables, learning rules, or deployment infrastructures while authorship remains external. To make this relocation explicit, this Perspective introduces a lifecycle map that locates where normative "whistles" enter AI systems from training through deployment, and proposes a constructive design pattern—the separation of an explicit Safety Envelope from an internal Adaptive Space—that preserves robust control while making normative authorship auditable and governable.

The Seduction of Internalization

A dominant paradigm has shaped the development of artificial agency. Under what might be called the *Optimizationist Programme*, artificial systems are framed as engines for objective-driven optimization. Whether trained through supervised learning or reinforcement learning, the system's purpose is defined externally: minimize a loss, maximize an expected return, converge until improvement stalls. The stop rule—"when enough is enough"—is imposed from outside [1–3].

In response to the brittleness of this arrangement, recent work in bio-inspired and autonomous AI has pursued a natural repair move: internalization. If the agent is given internal variables that encode sufficiency—needs, health, integrity, homeostasis—perhaps the referee can be dismissed [4–7]. Regulation replaces supervision. The appeal is immediate: smoother signals, greater robustness—and the tempting story that internal regulation implies intrinsic normativity.

This Perspective argues that this narrative is often mistaken. Often the constraint is not eliminated but relocated: the whistle becomes an internal reference the agent is trained to track. Replacing a referee's whistle with a thermostat dial does not remove the rule that governs the game; it changes where that rule is implemented. The thermostat regulates continuously and efficiently, but it does not decide what temperature ought to be maintained, nor why deviation should matter.

The contributions of this Perspective are threefold: (i) the *Rule-Relocation* lens, which distinguishes the location from the origin of normativity in artificial agents; (ii) a lifecycle map locating where normative constraints are inserted across training, architecture, meta-learning, and deployment;

and (iii) a constructive design pattern—*Safety Envelope + Adaptive Space*—that makes governance-authored constraints explicit while allowing rich internal regulation to develop within them.

This diagnosis does not reject internal regulation as engineering practice, but cautions against over-interpreting it as self-authored value. By making the whistle explicit rather than mythical, it becomes possible to reason more clearly about autonomy claims, design safer systems, and avoid mistaking disciplined control for intrinsic normativity.

1 The Rule-Relocation Problem

Bio-inspired regulation is often presented as a repair move for the Optimizationist Programme: internalize sufficiency and the system may appear to outgrow the need for an external referee. The claim here is diagnostic. In many contemporary designs, the normative constraint is not removed but *relocated*: the referee's whistle is compiled into an internal reference the agent is trained to track. This relocation can substantially improve learning dynamics and robustness without resolving where the constraint originates.

The whistle moves

In standard optimization, the whistle is overtly external: a loss function, reward definition, termination rule, or overseer decides what counts as success (for early stopping as an archetypal external predicate, see [1, 8]). In ostensibly self-regulating architectures, the whistle is often internalized as a regulatory target—health variables, drives, preferred-state distributions, or viability bounds [5, 6].

Rule-Relocation names the move from an external constraint to an internal target *without a corresponding shift in authorship*. Reference values, admissible regions, and violation semantics remain designer- or governance-specified, even when they are enforced by internal feedback loops [9].

The critical distinction is between *location* and *origin*. Internal variables can obscure authorship rather than dissolve it: a system may look self-directed while merely tracking imported references. Regulation answers how to remain within a boundary, not why that boundary holds nor whether it can be revised (contrast fixed setpoint control with biological regulation debates in [10, 11]). Making this distinction explicit prevents a reversal of explanation—inferring authorship from competence.

The claim is not that external authorship is always human, but that it is extra-agentic unless the system can revise what counts as binding constraint through its own continued organization and interaction.

A minimal formal anchor

Formal setup. Consider an agent interacting with an environment modeled as a Markov Decision Process with state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $P(s_{t+1}|s_t, a_t)$, and discount factor $\gamma \in (0, 1)$. Let $s_t \in \mathcal{S}$ denote the internal state at time t, $s^* \in \mathcal{S}$ a fixed reference state, and $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ a distance metric. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps states to distributions over actions.

Proposition 1 (Drive-reduction equivalence). Under a shaped reward $r_t = D(s_t, s^*) - D(s_{t+1}, s^*)$ and fixed reference s^*, maximizing expected drive reduction is equivalent to minimizing

expected distance to reference:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \big(D(s_t, s^*) - D(s_{t+1}, s^*)\big)\right] \iff \min_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t D(s_t, s^*)\right],$$

where $\mathbb{E}_{\pi}[\cdot]$ denotes expectation over trajectories $\tau \sim \pi$.

Proof sketch. Write $D_t := D(s_t, s^*)$. Then

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t (D_t - D_{t+1}) &= \sum_{t=0}^{\infty} \gamma^t D_t - \sum_{t=0}^{\infty} \gamma^t D_{t+1} \\
&= \sum_{t=0}^{\infty} \gamma^t D_t - \frac{1}{\gamma} \sum_{t=1}^{\infty} \gamma^t D_t \\
&= D_0 + \left(1 - \frac{1}{\gamma}\right) \sum_{t=1}^{\infty} \gamma^t D_t \\
&= D_0 - \frac{1-\gamma}{\gamma} \sum_{t=1}^{\infty} \gamma^t D_t.
\end{aligned}$$

Assuming D_t is bounded (or $\mathbb{E}_{\pi}[\sum_t \gamma^t D_t] < \infty$), maximizing the left-hand side is therefore equivalent to minimizing $\mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t D_t]$, since D_0 is fixed by the initial state distribution and the remaining term is a negative constant multiple of the discounted distance.

Interpretation. This equivalence holds when the reference s^* is fixed and externally specified; it is not a claim about all forms of adaptive regulation. Its purpose is diagnostic: when the reference is designer-authored and immutable, the "intrinsic" drive is an external objective implemented in internal form [12].

Canonical example: fixed priors and fixed setpoints

Relocation is clearest when preferences are specified as fixed setpoints (homeostatic reinforcement learning) [12] or fixed prior preferences (many Active Inference implementations) [14–16]. What changes is the engineering surface: dense signals replace sparse rewards, control becomes smoother, and anticipatory behavior can emerge [12, 13]. What does not change is authorship. The admissible region—safe set, viability kernel, or preferred-state manifold—remains specified, tuned, or ratified outside the agent [17, 18].

This is not a criticism of regulation as such. Relocation can be an essential safety pattern. The conceptual error is interpretive: treating fixed internal references as evidence that the agent has generated its own normativity (in the stronger, autopoietic sense) [10, 19]. Many systems can learn *how* to satisfy a constraint without being able to revise *what* counts as a constraint.

Why this matters

Relocation is easy to miss precisely because it works. Improved regulation supports robust control under imported norms and can mimic features associated with autonomy. The distinction motivates two constructive tools. Box 1 locates where whistles can reside across the system lifecycle, and

Section 4 separates what must remain governance-authored from what can safely adapt within those boundaries.

Box 1. Where the Whistle Lives: A Lifecycle Map of Rule-Relocation

This map is a portable diagnostic for locating where normative constraints ("the whistle") enter an AI system across its lifecycle—and for blocking the inference that internalization implies self-authorship.

1. **Training-time relocation.** Constraints arise from data support, optimization dynamics, or stopping rules. Agents "avoid failure" because certain trajectories are unrecoverable or unlearnable, not because failure carries intrinsic stake.
2. **Objective-time relocation.** Norms are written into loss or reward functions (penalties, side-constraints). The whistle is explicit but scalar, and therefore tradable.
3. **Architectural relocation.** Norms are compiled into internal state variables or feedback loops (health, integrity, preferred ranges). Regulation is internal; reference values and violation semantics remain designer-authored.
4. **Meta-learning relocation.** Norms are embedded in learning rules via an outer-loop objective (e.g., learned safety critics, learned constraint models, or meta-learned update rules). Behavior may appear self-directed, but success is still defined externally.
5. **Deployment-time relocation.** Norms are enforced during operation through monitoring, filters, overrides, or human governance. Even with internalized regulation, this layer often remains decisive.

Use. Internal regulation at one stage does not erase externally authored constraints at others. The map prevents cherry-picking and makes normative authorship legible across the system lifecycle.

2 Why Regulation Is Mistaken for Autonomy

Rule-relocation persists partly because regulation is impressive. Systems that keep internal variables within bounds can look agentic: they persist, self-correct, and remain stable under perturbation. In both biological discourse and bio-inspired AI, these surface features are often treated as evidence that a system has begun to author its own norms [5, 6].

Robust regulation explains *competence*, not *authorship*.

A controller can satisfy a constraint without having any authority over what counts as a constraint. Feedback and predictive regulation answer *how* to stay within a boundary; they do not answer *why that boundary holds*, nor whether it can be revised. A thermostat maintains temperature and a cruise controller maintains speed; neither authors the value it tracks [20]. Adding predictive depth can improve control without, by itself, relocating the origin of the constraint [11]. Regulation can therefore amplify the appearance of autonomy while leaving normative authorship unchanged [10].

Biological analogies make this confusion especially tempting. In living systems, viability is constitutive: failure to regulate is not merely degraded performance but loss of organization. The organism persists or disintegrates as a consequence of its own dynamics, so regulation, normativity, and survival are tightly coupled [21].

Contemporary AI lacks this coupling. What is called "death"—episode termination, reset, penalty, or shutdown—is typically a designer-authored event, not structural disintegration of the underlying system [22]. Failure can be reset or retrained; learning failure rarely entails disintegration. Internal "health" variables and termination conditions therefore function as simulated stakes: the system may learn *which* states trigger intervention, without any mechanism by which those boundaries become grounded in its own continued existence [9, 23].

The consequence is practical as well as conceptual. When engineered stakes are mistaken for intrinsic concern, governance authorship is obscured: norms appear to come from the agent rather than from design and deployment choices [24, 25]. And if normativity is assumed to be intrinsic, explicit oversight and auditing can be prematurely relaxed.

Recognizing this gap does not require metaphysical pessimism about artificial agency. It requires conceptual hygiene. Regulation may be necessary for autonomy, but it is not sufficient. The decisive question is not whether a system regulates itself, but whether the constraints it regulates are constitutive of its continued existence or compiled representations of external norms. This distinction motivates the graded analysis that follows (Section 3).

3 Degrees of Internalization

Rule-Relocation is not a binary verdict. Bio-inspired systems differ in how deeply regulation is embedded, how flexibly it adapts, and how tightly it is grounded in interaction history. The caution is narrower: increasing regulatory sophistication does not, by itself, relocate the origin of normativity.

3.1 Level I: Fixed setpoints (clear relocation)

The clearest case is regulation around fixed, designer-specified reference values: battery thresholds, integrity bounds, preferred internal states, or prior preferences. Dense internal signals replace sparse rewards, smoothing control and reducing brittle failure. Authorship, however, is unambiguous. The agent regulates what the designer chose to regulate, and the semantics of violation—reset, penalty, termination, or override—remain externally specified. The whistle is inside the architecture; its pitch is not.

3.2 Level II: Learned setpoints and policies (sophisticated relocation)

At the next level, systems learn regulatory targets, policies, or update rules from data and experience. Setpoints may be context-sensitive; objectives may be inferred rather than hard-coded; meta-learning may shape adaptation over time. But learning proceeds under an outer-loop criterion that fixes what counts as success, safety, or viability. The agent may learn *how* to regulate while remaining unable to contest *why* a particular constraint is binding. The referee is harder to see, not less decisive.

3.3 Level III: Predictive and allostatic regulation (degrees of internalization)

More recent architectures emphasize predictive, multi-timescale regulation over reactive error correction. Anticipatory control enables action before thresholds are breached, trades off competing demands, and improves performance under non-stationarity. In this sense, allostatic regulation deepens internalization.

However, increased predictive depth improves control competence without, by itself, relocating the origin of the constraint. Unless the system can revise what counts as viable through sustained interaction—rather than merely optimizing predictions within a fixed evaluative frame—normative authorship remains external. The whistle becomes a prediction error the system is permanently tasked to suppress [26].

Audit axes for internalization

To prevent these distinctions from collapsing into narrative appeal, three audit axes are useful:

- **Setpoint mutability.** Are targets fixed, contextually modulated, learned, or revisable through interaction? Can the agent alter what counts as "enough," or only how to achieve it?
- **Grounding in interaction history.** Are constraints grounded in sustained interaction, or inferred under fixed outer-loop objectives? Does the agent update a model of viability, or only policy within it?
- **Causal insulation of constraints.** Can the agent influence the definition of its own constraints, or are boundaries insulated by design? Can self-modification affect norms, or only behavior under them?

Internalization admits degrees; disappearance does not. Recognizing this preserves the value of bio-inspired regulation—robustness, flexibility, and safety—while keeping governance and authorship explicit. It also prepares the ground for the illustrative case that follows and the constructive design pattern in Section 4.

Contemporary examples

These distinctions apply directly to current systems. Recent empirical work demonstrates that behavioral norms in deployed LLMs—such as calibrated uncertainty expression or refusal patterns—can be imposed by external decision rules operating on internal model states, without modifying the underlying model [27]. Similarly, Constitutional AI makes normative commitments explicit as an external set of principles and enforces them via post-training feedback (and associated deployment practices), rather than treating them as self-authored, revisable commitments of the model [30]. Both examples instantiate Rule-Relocation: normativity is enacted at system boundaries (monitoring, routing, and filtering layers) rather than internalized as revisable commitments. The agent-like core may display rich self-description and apparent flexibility, yet decisive normative constraints remain externally specified and are not renegotiable within the interaction.

4 Safety Envelope and Adaptive Space

If the Rule-Relocation Problem marks an interpretive error—reading internal regulation as self-authored normativity—the constructive response is not less regulation, but clearer authorship. This Perspective proposes a simple design pattern: a *Safety Envelope* paired with an *Adaptive Space*. The envelope makes safety constitutive by defining non-negotiable boundaries on permissible trajectories; within those boundaries, the adaptive space permits exploration, planning, and learning.

Safety as a constitutive constraint

Penalties and auxiliary objectives invite ambiguity. They render safety tradable, encourage proxy optimization, and obscure where normative commitments are enforced [28]. A Safety Envelope avoids this by treating safety as a proscriptive constraint—a set-invariance commitment—so violation is not merely costly but disallowed by design [29]. The aim is not to simulate concern, but to encode ethical, legal, or safety-critical commitments that must remain insulated from the agent's own optimization dynamics [30, 31].

Outer loop and inner loop

The pattern separates normative labor:

- **Outer Loop (Safety Envelope).** Defines viability conditions, monitors proximity to boundaries, and intervenes to prevent violation. Its logic is conservative, auditable, and hard for the agent to modify.
- **Inner Loop (Adaptive Space).** Learns and plans within the envelope. Here autonomy is permitted: the agent can be exploratory, predictive, and opportunistic, so long as proposed actions remain inside the envelope.

This separation subsumes many implementations—constrained reinforcement learning, shielding, barrier functions, runtime monitors—without depending on any one of them [29, 32, 33]. What matters is not the mechanism but the partition: constitutive constraints outside; adaptive objectives inside.

Why explicit envelopes reduce failure modes

Relocating safety into internal drives or dense shaping signals often produces two familiar pathologies. First, *rule-hiding*: when constraints are implicit in learned representations, they can become opaque, spurious, and difficult to audit [35]. Second, *meditative-agent failure*: when minimizing internal error is the mandate, inactivity can become the safest policy [28].

Explicit envelopes blunt both. They keep constraints legible and structural rather than statistical, and they decouple safe operation from the agent's incentive to collapse uncertainty. This separation matters most near irreversible "sink" states, where intrinsic objectives and exploration bonuses can fail to protect viability [36].

Instrumental coupling without normativity illusion

Inside an envelope, persistence is instrumentally necessary: continued operation is a precondition for action. That coupling is desirable. What should be resisted is the interpretive leap from engineered necessity to intrinsic concern. The agent persists because persistence enables task pursuit, not because it has authored a value. Instrumental self-preservation can be bounded and audited; claims of intrinsic normativity cannot [37].

Governance and authorship

Because the envelope encodes non-negotiable commitments, its authorship must remain explicit. Changes to the envelope should occur through authorized design and governance processes, not through the agent's learning dynamics. This explicitness makes safety claims inspectable: one can point to the envelope, test it, and verify whether it persists under distribution shift, reward thinning, or self-modification [38–40].

Positioning

The Safety Envelope + Adaptive Space pattern does not reject bio-inspired regulation; it gives it a place. Predictive control and multi-timescale regulation can flourish inside the envelope without being burdened with claims of self-authored normativity, for example through hierarchical generative control architectures [34]. By making the whistle explicit, the pattern avoids both romanticization and denial: some norms must remain externally authored and governed, while internal regulation can grow where it reliably delivers robustness and competence.

Implementation challenges and trade-offs

Specifying safety envelopes for complex, real-world norms presents significant engineering challenges. In high-dimensional settings, the boundary conditions that implement a norm may be difficult to hard-code and may instead be specified as externally authored rules or principles that are operationalized through feedback and deployment mechanisms [30]. Envelope design also involves a fundamental trade-off: overly conservative boundaries can yield brittle or overly cautious agents, while overly permissive ones may fail to prevent harm [17, 32]. Distribution shift and novel situations can challenge pre-specified envelopes, motivating mechanisms for safe adaptation and, when necessary, authorized revision [39]. When boundaries are refined through interaction or feedback, maintaining explicit authorship and audit trails becomes more difficult but remains essential. These challenges motivate future work on envelope specification methods, verification under distribution shift, and governance protocols for authorized constraint revision. The pattern proposed here is a design principle, not a complete implementation; its value lies in making normative authorship legible and separable from adaptive competence.

The next box distills this pattern into a compact audit checklist aimed at preventing Rule-Relocation from collapsing into Rule-Hiding.

Box 2. Auditing Rule-Relocation

This checklist distinguishes *Rule-Relocation* from *Rule-Hiding* by making the location, persistence, and authorship of normative constraints explicit.

1. **Where does the whistle live?** At which lifecycle stages—training, objective, architecture, meta-learning, deployment—are constraints enforced?
2. **Can the agent influence the envelope?** Is there any causal pathway by which the agent can modify or redefine the constraint itself, rather than merely act within it?
3. **Is survival instrumental or terminal?** Does satisfying safety conditions enable continued task pursuit, or does the agent collapse into inactivity once internal metrics are satisfied?
4. **Is violation structural or scalar?** Are boundary breaches non-negotiable (override, termination), or are they penalties that can be outweighed by reward?
5. **Do constraints persist under self-modification?** If the system updates itself, trains successors, or thins task reward, do the same commitments remain unless explicitly altered through authorized governance?

Use. Systems that cannot answer these questions clearly are likely relocating normativity without making its authorship legible.

Make the Whistle Explicit

The central claim of this Perspective is simple: in much bio-inspired AI, the whistle rarely disappears—it moves. Stopping conditions, safety rules, and normative commitments are relocated into internal variables, setpoints, priors, and regulatory loops. When those references remain fixed and designer-authored, what changes is the *location* of normativity, not its *origin*.

This diagnosis does not reject internal regulation as engineering practice, but cautions against over-interpreting it as self-authored value. Homeostatic and allostatic mechanisms can improve robustness, flexibility, and control; the mistake is to treat those gains as evidence that a system has authored its own ends.

Making the whistle explicit is a safety issue, not merely a conceptual one. When normative constraints are compiled into shaping terms, learned representations, or anthropomorphic narratives, their authorship becomes obscured. Hidden normativity is riskier than explicit governance: it is harder to audit, harder to revise, and easier to misattribute to the system itself.

The alternative advanced here is a clearer partition of normative labor. A *Safety Envelope* makes non-negotiable commitments structural, explicit, and governance-authored. Within that envelope, an *Adaptive Space* permits rich internal regulation, predictive control, and learning without importing philosophical claims they cannot support. The lifecycle map specifies where constraints enter across the system lifecycle, while the audit checklist tests whether those constraints persist under adaptation and self-modification.

The most promising path for bio-inspired AI is therefore neither the romanticization of simulated needs nor the denial of internal regulation. It is to keep some norms visibly authored, governed, and enforceable, while allowing internal regulation to grow where it reliably delivers competence. By

making the whistle visible rather than mythical, we enable safer systems and clearer claims about autonomy.

References

[1] Lutz Prechelt. Early stopping - but when? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pages 55–69. Springer, Berlin, Heidelberg, 1998. doi: 10.1007/3-540-49430-8_3.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. Preprint at `https://arxiv.org/abs/1412.6980`, 2014.

[3] Emre O. Neftci and Bruno B. Averbeck. Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, 1(3):133–143, 2019. doi: 10.1038/s42256-019-0025-4.

[4] Lisa Feldman Barrett. The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017. doi: 10.1093/scan/nsw154.

[5] Kingson Man and Antonio Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10):446–452, 2019. doi: 10.1038/s42256-019-0103-7.

[6] Matthew Sims. Self-concern across scales: A biologically inspired direction for embodied artificial intelligence. *Frontiers in Neurorobotics*, 16:857614, 2022. doi: 10.3389/fnbot.2022.857614.

[7] Nature Machine Intelligence. A soft touch for robots. *Nature Machine Intelligence*, 4(5):415, 2022. doi: 10.1038/s42256-022-00496-2.

[8] Lutz Prechelt. Automatic early stopping using cross validation: Quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998. doi: 10.1016/S0893-6080(98)00010-0.

[9] Tomas Veloz. Toward aitiopoietic cognition: bridging the evolutionary divide between biological and machine-learned causal systems. *Frontiers in Cognition*, 4:1618381, 2025. doi: 10.3389/fcogn.2025.1618381.

[10] Tom Froese and Tom Ziemke. Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4):466–500, 2009. doi: 10.1016/j.artint.2008.12.001.

[11] Peter Sterling. Allostasis: A model of predictive regulation. *Physiology & Behavior*, 106(1):5–15, 2012. doi: 10.1016/j.physbeh.2011.06.004.

[12] Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife*, 3:e04811, 2014. doi: 10.7554/eLife.04811.

[13] Hugo Laurençon, Charbel-Raphaël Ségerie, Johann Lussange, and Boris S. Gutkin. Continuous homeostatic reinforcement learning for self-regulated autonomous agents. Preprint at `https://arxiv.org/abs/2109.06580`, 2021.

[14] Pablo Lanillos, Cristian Meo, Corrado Pezzato, Ajith Anil Meera, Mohamed Baioumy, Wataru Ohata, Alexander Tschantz, Beren Millidge, Martijn Wisse, Christopher L. Buckley, and Jun

Tani. Active inference in robotics and artificial agents: Survey and challenges. Preprint at https://arxiv.org/abs/2112.01871, 2021.

[15] Corrado Pezzato, Riccardo Ferrari, and Carlos Hernandez. A novel adaptive controller for robot manipulators based on active inference. Preprint at https://arxiv.org/abs/1909.12768, 2019.

[16] Pietro Mazzaglia, Tim Verbelen, Ozan Çatal, and Bart Dhoedt. The free energy principle for perception and action: A deep learning perspective. *Entropy*, 24(2):301, 2022. doi: 10.3390/e24020301.

[17] Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9797–9807. PMLR, 2020. URL https://proceedings.mlr.press/v119/wachi20a.html.

[18] Alexander Gerdt, Nikolai Botkin, Johannes Diepolder, Varvara Turova, and Florian Holzapfel. 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, 2020. URL https://ceur-ws.org/Vol-2783/paper06.pdf.

[19] Francesco Bianchini. Autopoiesis of the artificial: from systems to cognition. *BioSystems*, 230: 104936, 2023. doi: 10.1016/j.biosystems.2023.104936.

[20] Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow. Behavior, purpose and teleology. *Philosophy of Science*, 10(1):18–24, 1943. doi: 10.1086/286788.

[21] Francisco G. Varela, Humberto R. Maturana, and Ricardo Uribe. Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5(4):187–196, 1974. doi: 10.1016/0303-2647(74)90031-8.

[22] Xabier E. Barandiaran. Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, 36(3):409–430, 2017. doi: 10.1007/s11245-016-9365-4.

[23] Marcin Korecki, Cesare Carissimo, and Tanner Lund. artificial death: learning from stories of failure. In *Proceedings of the 2023 Conference on Artificial Life*, volume 35, pages 41–49. MIT Press, 2023. doi: 10.1162/isal_a_00633.

[24] David Watson. The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440, 2019. doi: 10.1007/s11023-019-09506-6.

[25] Nicholas Barrow. Anthropomorphism and ai hype. *AI and Ethics*, 4(3):707–711, 2024. doi: 10.1007/s43681-024-00454-1.

[26] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2 (4):210–219, 2020. doi: 10.1038/s42256-020-0170-9.

[27] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu,

Lukas W. Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231, 2025. doi: 10.1038/s42256-024-00976-7.

[28] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. Preprint at `https://arxiv.org/abs/1606.06565`, 2016.

[29] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control barrier functions: Theory and applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431. IEEE, 2019. doi: 10.23919/ECC.2019.8796030.

[30] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. Preprint at `https://arxiv.org/abs/2212.08073`, 2022.

[31] Gianluca Baldassarre, Richard J. Duro, Emilio Cartoni, Mehdi Khamassi, Alejandro Romero, and Vieri Giuliano Santucci. A formalisation of the purpose framework: the autonomy-alignment problem in open-ended learning robots. Preprint at `https://arxiv.org/abs/2403.02514`, 2025.

[32] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 22–31. PMLR, 2017. URL `https://proceedings.mlr.press/v70/achiam17a.html`.

[33] Richard Cheng, Gábor Orosz, Richard M. Murray, and Joel W. Burdick. 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, 2019. doi: 10.1609/aaai.v33i01.33013387.

[34] Kai Yuan, Noor Sajid, Karl Friston, and Zhibin Li. Hierarchical generative modelling for autonomous robots. *Nature Machine Intelligence*, 5:1402–1414, 2023. doi: 10.1038/s42256-023-00752-z.

[35] Maximilian Dreyer, Jim Berend, Tobias Labarta, Johanna Vielhaben, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Mechanistic understanding and validation of large AI models with SemanticLens. *Nature Machine Intelligence*, 7:1572–1585, 2025. doi: 10.1038/s42256-025-01084-w.

[36] Thomas A. Berrueta, Allison Pinosky, and Todd D. Murphey. Maximum diffusion reinforcement learning. *Nature Machine Intelligence*, 6:504–514, 2024. doi: 10.1038/s42256-024-00829-3.

[37] Stephen M. Omohundro. The basic AI drives. In Pei Wang, Ben Goertzel, and Stan Franklin,

editors, *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pages 483–493. IOS Press, 2008.

[38] Anqi Li, Dipendra Misra, Andrey Kolobov, and Ching-An Cheng. Survival instinct in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, Red Hook, NY, 2023. Curran Associates, Inc. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/c3e969ea20542a6a11e6caeac736a0b9-Paper-Conference.pdf`.

[39] Mayank Kejriwal, Eric Kildebeck, Robert Steininger, and Abhinav Shrivastava. Challenges, evaluation and opportunities for open-world learning. *Nature Machine Intelligence*, 6:580–588, 2024. doi: 10.1038/s42256-024-00852-4.

[40] The Editors. Seeking clarity rather than strong opinions on intelligence. *Nature Machine Intelligence*, 6:1408, 2024. doi: 10.1038/s42256-024-00968-7.