frontiers | Frontiers in Cognition

TYPE Conceptual Analysis
PUBLISHED 17 July 2025
DOI 10.3389/fcogn.2025.1618381

OPEN ACCESS

EDITED BY
Andrew Tolmie,
University College London, United Kingdom

REVIEWED BY
Sheila L. Macrine,
University of Massachusetts Dartmouth, United States
Lucas Fucci Amato,
University of São Paulo, Brazil

*CORRESPONDENCE
Tomas Veloz
✉ tomas.velozg@utem.cl

RECEIVED 25 April 2025
ACCEPTED 18 June 2025
PUBLISHED 17 July 2025

CITATION
Veloz T (2025) Toward aitiopoietic cognition: bridging the evolutionary divide between biological and machine-learned causal systems. *Front. Cognit.* 4:1618381. doi: 10.3389/fcogn.2025.1618381

Toward aitiopoietic cognition: bridging the evolutionary divide between biological and machine-learned causal systems

Tomas Veloz*

Departamento de Matemáticas, Universidad Tecnológica Metropolitana, Santiago, Chile

We examine and compare autopoietic systems (biological organisms) and machine learning systems (MLSs) highlighting crucial differences in how causal reasoning emerges and operates. Despite superficial functional similarities in behavior and cognitive abilities, we identify profound structural differences in how causality is operationalized, physically embodied, and epistemologically grounded. In autopoietic systems, causal reasoning is intrinsically tied to self-maintenance processes across multiple organizational levels, with goals emerging from survival imperatives. In contrast, MLSs implement causality through statistical optimization with externally imposed objectives, lacking the material self-reorganization that drives biological causal advancement. We introduce the concept of "aitiopoietic cognition"—from Greek "aitia" (cause) and "poiesis" (creation)—as a framework where causal understanding emerges directly from a system's self-constituting processes. Through analyzing convergence pathways including evolutionary algorithms, material intelligence, homeostatic regulation, and multi-scale integration, we propose a research program aimed at bridging this evolutionary divide. Such integration could lead to artificial systems with genuine intrinsic goals and materially grounded causal understanding, potentially transforming our approach to artificial intelligence and deepening our comprehension of biological cognition.

KEYWORDS

artificial intelligence, emergence, causal reasoning, autopoieisis, metasystem transitions, embodied cognition, synthetic biology

1 Introduction

Machine learning systems (MLSs) are rapidly increasing their influence in our lives, transforming sectors from healthcare to entertainment (Marcus and Davis, 2019; LeCun et al., 2015), and where substantial investments are flowing into their sophistication (Maslej et al., 2025), there is an increasing need to understand the fundamental differences between "us and them" (Bengio et al., 2024). For clarity and analytical precision, we will refer to "them" as MLSs, acknowledging their primary mechanism of development and adaptation.

Considering the perspective of a child first learning to distinguish entities in a computer-interface level, or for an alien visiting our planet, there is no major difference between humans and MLSs. Both can read, understand text and images, type and draw, speak, and engage in a stream of complex actions including planning abilities, communication skills, and even abstract reasoning about concepts, causal relationships, and reflexive understanding of self and others. This similarity is recognized as well in

robotic interfaces that allow for physical interaction and movement (Moro et al., 2019; Manzi et al., 2020). This seemingly remarkable similarity is often explained through the lens of computational functionalism (Putnam, 1967; Chalmers, 1996), which posits that mental states are defined by their functional roles rather than their physical substrate. Under this view, if MLSs functionally replicate human cognitive processes, even from a completely different substrate, they can be considered fundamentally equivalent—and hence subjected to comparison at the agential level (Goertzel, 2007).

1.1 The "equivalence hypothesis" for testing causal cognition in human and machines

By looking more closely at how we compare ourselves and machines, we arrive at the strong influence that has played the Turing test, which is assumed to be known by the reader. While it aims at testing thinking, it more precisely tests the ability to engage in a conversation using previously learned information, and hence it does not test thinking directly, but learning and the ability to express such learning (Moor, 1976). In this article we do not want to dig into the definitions of learning, thinking and intelligence and how that impacts our understanding of artificial intelligence (see Wang, 2019 for such analysis). Instead, we want to explore the consequences of comparing the causal cognitive abilities of a machine and a human using input-output architectures. Mainstream cognitive science implicitly assumes that these architectures "mean the same" for both humans and machines. In fact, the input-output architecture is, as well, generally believed to be responsible for the learning process, in an equivalent way, for both humans and MLSs. The latter is accepted and well-justified by a large number of successful scientific programs in a variety of fields including various branches of psychology and cognitive science (Anderson, 2007), linguistics (Fodor, 1975; Chomsky, 1986; Pinker, 1994), computer science (Russell and Norvig, 2020), ethology (Lorenz, 1981), and others (Clark, 2001). These have shown that the design of processes and experiments based on input-output architectures have a high inductive power and allow to explain how learning, adaptation, and the development of increasingly complex responses (behaviors) to environmental challenges can be generated/stopped or enhanced/inhibited. Therefore, it is assumed that the way in which the "processing unit" that transforms input to output, i.e. the "black-box" for MLSs or "the mind" for humans, has no particular difference for what concerns defining cognition (Chalmers, 1996; Clark and Chalmers, 1998).

However, the above implicit equivalence assumption masks profound differences in how the inner workings of the material implementations of the input and output, and more crucially of them together with the "processing unit" forming a full system, shapes the existential, developmental and evolutionary features of cognition in MLSs and humans (Deacon, 2011).

At the existential level, cognition in Humans and other biological organisms is implemented within autopoietic systems—self-creating and self-maintaining entities that constantly regenerate their components through metabolic processes that harness energy from its environment (Boden, 1999; Maturana and Varela, 2012; Thompson, 2007). In contrast, cognition in MLSs is implemented in physically static machines whose embodiment remains unchanged. At the developmental level autopoietic systems develop through multi-level structures based on cells made of molecular networks (Fields and Levin, 2022; Witkowski et al., 2023), each level having its own sense of self and its own competences resembling cognitive abilities aligned to their particular physical instantiation (which might or might not implement universal Turing computation), while for MLSs their development is based on external assemblage without multi-level structures and no internalized sense of self, and a "single level of intelligence" evolves through algorithmic adjustments and data-driven feedback in a universal Turing machine setting.

At the evolutionary level, these differences amplify what cognition means at each substrate. Autopoietic systems have evolved through natural selection operating on genetic variations across billions of years, with multiple major transitions creating hierarchical levels of organization (Smith and Szathmáry, 1995; Szathmáry, 2015). This evolutionary process has resulted in systems where purposeful behavior emerges from the intricate interplay between material constraints and informational dynamics at multiple scales (West et al., 2015; Heylighen, 2023). In stark contrast, MLSs "evolve" through directed human engineering, following developmental trajectories on a fixed Von Neumann architecture, optimized for specific). purposes involving performance, economic cost, size, and computation speed, rather than survival in open-ended environments (Stanley and Lehman, 2015). Their evolutionary trajectory lacks the self-organized complexity and emergent properties characteristic of biological evolution, instead following design principles imposed externally by human developers with predetermined objectives (Lake et al., 2017).

1.2 Goals as a criteria for comparing causal cognition

Instead of focusing on performance to test causal cognition, we can focus on "goal-directedness," as goals reflect the source of actions in the processing unit (Heylighen, 2023). Following this idea, goals serve as a crucial pivot point for comparing the causal cognition between humans and machines (Deacon, 2011). In autopoietic systems, goals are defined by the relation between the imperative for the system's physical existence through self-preservation and the ways available to interact with its environment (Kolchinsky and Wolpert, 2018). For MLSs, goals are externally defined optimization targets disconnected from any necessity of material self-preservation. This difference in the origin and nature of goals highlights an important gap in how their causal cognition operates. This aspect raises issues regarding the comparison of other significant aspects of cognition such as intelligence and adaptation (see Stano et al., 2023, and other articles in that special issue).

This paper examines the fundamental characteristics of autopoiesis and machine learning, analyzes their key differences using goals as a reference concept that concern causal cognition, and explores potential convergence in futuristic systems that

integrate their diverse "cognitive scaffoldings" (Ziemke et al., 2004). Finally we outline a path toward bridging this evolutionary divide (Witkowski et al., 2023; Seth, 2021), by proposing "aitiopoietic cognition"—from Greek "aitia" (cause) and "poiesis" (creation)—as a framework where causal understanding emerges directly from a system's self-constituting processes, creating a recursive relationship between physical organization and causal reasoning.

2 Autopoietic systems: from survival to goals

Autopoietic systems, introduced by biologists Humberto Maturana and Francisco Varela in the 1970s, are defined as networks of processes that produce the components necessary for their continued existence and boundary maintenance. This concept provides a cybernetic-inspired framework for understanding biological autonomy (Maturana and Varela, 2012). Unlike mechanistic or vitalistic accounts of life, autopoiesis offers a naturalistic perspective that emphasizes the dynamic, process-oriented nature of living systems (Weber and Varela, 2002).

2.1 The inner-outer structure

From a biochemical perspective, autopoietic systems operate through metabolic networks that continuously transform matter and energy to regenerate their components while maintaining organizational stability. This self-production occurs through thermodynamically open processes that sustain the system far from equilibrium (Moreno and Mossio, 2015). The continuous production of a semi-permeable boundary distinguishes the system from its environment while regulating internal processes and external exchanges, creating a fundamental inside-outside asymmetry crucial for biological autonomy (Luisi, 2003).

A defining characteristic of autopoietic systems lies in their hierarchical organization across spatial and temporal scales. This multi-level architecture enables nested autonomy: molecular networks maintain metabolic closure, cells exhibit decision-making via signaling pathways (Gao et al., 2023), and tissues coordinate via biophysical feedback (Forgacs and Newman, 2005). Each level sustains operational closure while contributing to higher-order autopoiesis, exemplifying Varela's concept of "autonomous identity at several levels" (Varela, 1979). This organization aligns with Salthe's (1985) hierarchical evolution framework and enables bidirectional causality: "downward causation" (Ellis, 2012), where higher levels modulate lower-level processes, and "upward constraints," where molecular dynamics limit higher-level possibilities (Kauffman, 1993; West-Eberhard, 2003).

The operational closure of autopoietic systems does not preclude dynamic engagement with environments; rather, it enables multi-level structural coupling—a process by which recurrent interactions trigger compensatory changes while preserving organizational coherence (Di Paolo, 2005; Moreno and Mossio, 2015). This coupling operates hierarchically: molecular networks couple with intracellular conditions, cells with tissue microenvironments, and organisms with ecological niches, each level maintaining its autonomy while contributing to the system's viability (Maturana and Varela, 2012; Salthe, 1985). Evolutionary pressures sculpt this hierarchy, favoring modularity for robustness (Wagner, 1996) and degeneracy (Edelman and Gally, 2001) to buffer against perturbations.

Crucially, coupling is asymmetrical and structurally determined: the environment does not dictate changes but perturbs the system, whose architecture—shaped by evolutionary and developmental history—filters which perturbations are salient (Barandiaran and Moreno, 2008; Juarrero, 1999). A cell's membrane receptors selectively respond to extracellular ligands while its metabolic state constrains receptor expression, illustrating how multi-level dependencies mediate environmental interactions (West-Eberhard, 2003; Huang, 2012; Rafelski and Theriot, 2024).

Thus, autopoietic systems enact their worlds through multi-scalar, history-laden interactions—a process where autonomy and dependency coexist, and every perturbation becomes an opportunity for meaning-making (Varela et al., 1991; Barandiaran et al., 2009).

2.2 Autopoietic cognition

The connection between autopoiesis and cognition emerges from the system's need to maintain itself through adaptive interactions with its environment. As Maturana and Varela provocatively stated, "living is knowing," suggesting that even basic autopoietic systems exhibit a primitive form of cognition through their selective environmental coupling (Thompson, 2007). This perspective reframes cognition not as information processing but as sense-making—transforming neutral environmental stimuli into meaningful distinctions relevant to continued existence (Di Paolo and Thompson, 2014).

The transition of autopoietic systems into sense-making adaptive agents hinges on their capacity to develop multi-level regulatory hierarchies that monitor and modulate viability conditions across scales (Di Paolo, 2005; Moreno and Mossio, 2015). From these processes goal-directedness becomes naturally linked to autopoietic organization. Autopoietic systems exhibit "purposive" behavior because their structure—forged through structural coupling (Juarrero, 1999)—embodies historical solutions to viability challenges (Deacon, 2011). For instance, slime molds optimize nutrient networks via self-organizing gradients (Nakagaki et al., 2000). This naturalized teleology (Weber and Varela, 2002) proposes a solution to the paradox of purpose as entities in the future influencing its past: goals are emergent properties of systems that recursively couple action to self-maintenance across scales (Barandiaran et al., 2009).

The formal representation of goals in autopoietic systems presents unique modeling challenges precisely because goals emerge from the system's organization rather than being explicitly encoded (Veloz, 2021). Several mathematical frameworks have been developed to capture this emergence, each highlighting different aspects of how purpose arises from process.

Dynamical systems theory provides the broadest framework, representing autopoietic goals as attractor states in state space that maintain viability amid perturbations (Heylighen, 2023; Kauffman,

1993). Crucially, these attractors shift based on the system's internal state, creating a landscape where goals are context-dependent rather than fixed. This adaptive landscape model has been formalized in work on viability boundaries and adaptive control (Barandiaran and Egbert, 2014), providing mathematical tools to analyze how autopoietic systems generate and modify goals in response to changing conditions.

More specific mathematical frameworks address different aspects of goal-directedness. Chemical Organization Theory (Dittrich and Speroni di Fenizio, 2007; Veloz and Razeto-Barry, 2017) models self-maintaining chemical networks where closure and self-production create stability conditions that serve as implicit goals. This approach enables rigorous analysis of how chemistry creates persistent identity—a precondition for purpose—by identifying organizational closure in reaction networks. Meanwhile, Free Energy Principle models (Friston, 2010; Priorelli et al., 2025) formalize how predictive regulation serves autopoietic maintenance, representing goals as probability distributions over viable states that systems act to maintain through active inference. This Bayesian approach provides a computational bridge between autopoietic goals and machine learning frameworks. Agent-based modeling has become increasingly important for modeling how goals emerge from simple behavioral mechanisms tied to viability constraints, demonstrating how selection pressures can drive the emergence of increasingly complex goal hierarchies through simulated evolution (Froese and Ziemke, 2009; Packard et al., 2019).

While the conceptualization of goals is thoroughly integrated with theoretical frameworks of autopoiesis, and the relationship between goal-directedness and biological purpose is well-established in philosophical terms, the formal mathematical representation of these concepts remains in its infancy. Current models capture aspects of emergent purpose but struggle to represent the full richness of biological goal-directedness—particularly the multi-scale competencies that comprise biological intelligence (Witkowski et al., 2023; Fields and Levin, 2022). These competencies, from basal cognition in single cells to the abstract reasoning of complex organisms, suggest multiple forms of intelligence operating across different scales and materialities, challenging our current modeling capabilities. As autopoietic theory continues to evolve, bridging the gap between rich theoretical accounts and precise formal representations remains a crucial frontier—one that will not only deepen our understanding of biological cognition but also provide insights for developing artificial systems with more naturalistic forms of goal-directedness (Veloz, 2021; Thompson, 2007; Di Paolo, 2005).

2.3 From autopoiesis to aitiopoiesis

The transition from autopoietic to aitiopoietic cognition, i.e., going from achieve material self-preservation to embody affordances through behaviors that resemble causal reasoning, represents a fundamental leap where systems transcend mere organizational closure to actively constitute causal knowledge through their very existence (Deacon, 2011). This transition becomes evident when examining how autopoietic systems generate what we term "agential causality"—causal understanding that emerges not from abstract computation but from the material processes of self-constitution and environmental coupling. Consider bacterial chemotaxis: the cell's sensory-motor apparatus doesn't simply detect gradients but constitutes a knowledge-generating system where the phosphorylation cascade dynamics simultaneously maintain cellular organization and create understanding about environmental cause-effect relationships (Davies and Levin, 2023; Levin, 2019). The bacteria's tumble-and-run behavior emerges from constitutional processes where causal learning and self-maintenance are inseparably intertwined—the system literally embodies its causal models through the recursive dynamics of its own material organization.

Recent advances in synthetic multicellularity illuminate this transition by revealing how collective systems can exhibit emergent aitiopoietic properties that exceed their individual components' autopoietic capabilities. Xenobots, constructed from amphibian skin and cardiac cells, demonstrate a primitive form of aitiopoietic cognition where collective behavior emerges from cellular self-organization without genetic programming or external control circuits (Moreno and Etxeberria, 2005; Newman and Bhat, 2009; Kriegman et al., 2020). These "living robots" navigate their environment through constitutive processes—their locomotion, object manipulation, and collective coordination arise from the same self-maintaining dynamics that preserve their multicellular integrity. Crucially, their behavioral competencies represent endogenous properties of agential materials rather than externally imposed algorithms, suggesting that aitiopoietic cognition scales naturally from autopoietic foundations when appropriate organizational architectures emerge. Similarly, Anthrobots self-assemble from human lung cells into motile spheroids with cilia-driven propulsion and tissue-repair capabilities, demonstrating how multicellular collectives can exhibit goal-directed behaviors that emerge from, rather than being programmed into, their constitutional dynamics (Solé et al., 2024).

The synthetic biology framework reveals aitiopoiesis as fundamentally involving multi-scale agency where causal competencies emerge through hierarchical coupling between different levels of organization (Solé et al., 2016). In organoid systems, individual cells contribute to tissue-level morphodynamic reasoning—the collective navigation of anatomical morphospace through perception-action loops that simultaneously maintain tissue architecture and generate knowledge about spatial relationships (Solé et al., 2024). This represents a form of "collective aitiopoietic cognition" where causal understanding emerges from the constitutional dynamics of cellular collectives, not from pre-programmed instructions. The tissue "learns" about its environment through the very processes that constitute its existence—growth gradients, mechanical forces, and bioelectrical patterns become both the medium of self-maintenance and the substrate of causal reasoning. Importantly, this scaling of aitiopoietic competency reveals a crucial principle: higher-order causal understanding doesn't reduce to lower-level mechanisms but emerges through what Levin terms "agential materials"—substrates with intrinsic competencies that can be guided through behavioral interventions rather than mechanical control.

The implications for artificial aitiopoietic systems are profound. Current synthetic approaches reveal that genuine

aitiopoiesis cannot be engineered through traditional top-down design but must emerge from substrates that exhibit "competency in transcriptional, anatomical, and physiological problem spaces" (Davies and Levin, 2023). The failure of purely computational approaches to achieve constitutional causality suggests that future aitiopoietic systems will require physical substrates where information processing affordances such as pattern recognition and causal reasoning as well as self-maintenaning affordances such as reparation and duplication are materially unified rather than functionally separated (Gill et al., 2025). This points toward a research program focused not on programming artificial agents but on cultivating synthetic systems where aitiopoietic cognition can emerge from the recursive dynamics of embodied self-organization, potentially through hybrid bio-synthetic architectures that combine the constitutional properties of living materials with the scalability of artificial substrates.

3 Machine learning systems: from goals to causality

Machine learning represents the current paradigm to artificial intelligence by allowing algorithms to improve through experience by learning from data (Russell and Norvig, 2020). Early machine learning focused on representational approaches and decision trees, but the field underwent a transformation with the advent of deep learning architectures that could automatically extract hierarchical features from raw data (LeCun et al., 2015). This evolution reflects a broader transition from engineering-centric to data-centric approaches, where system behavior emerges from statistical patterns rather than explicit design.

3.1 Machine learning instantiations of causal cognition

At its core, machine learning operates through statistical inference and optimization processes that adjust internal parameters to minimize prediction errors or maximize reward signals. MLSs operationalize causality through statistical associations rather than mechanistic or teleological reasoning. Rooted in pattern recognition, these systems optimize for predictive accuracy by minimizing loss functions (e.g., cross-entropy, mean squared error) that measure deviations from training data distributions (Goodfellow et al., 2016). According to Pearl's (2019) Causal Hierarchy, ML systems operates at Level 1 (observational inference), lacking capacity for intervention (Level 2) or counterfactual reasoning (Level 3). For instance, deep neural networks trained on medical datasets may correlate hospital beds with patient mortality without inferring beds as sites of treatment rather than causation (Geirhos et al., 2020). Such spurious correlations stem from ML's reliance on statistical shortcuts—surface features that maximize training accuracy but fail to capture causal invariance (Arjovsky et al., 2019).

This process can be formalized mathematically as finding a function f* such that prioritizes empirical risk minimization:

$$f^* = argmin_f E[L(f(x), y)], \quad (1)$$

where E means expected value, L quantifies prediction error over a dataset D. While effective for interpolating training distributions, this formulation conflates correlation with causation, as models lack mechanisms to distinguish confounding variables (Schölkopf, 2022). For example, ML systems trained on socioeconomic data often reproduce biased associations (e.g., race and loan default rates) due to dataset imbalances rather than causal relationships (Koh et al., 2021).

Recent critiques highlight how this statistical foundation limits ML's causal robustness. Adversarial attacks—minor input perturbations that deceive models (Szegedy et al., 2013)—expose the fragility of associational reasoning, while distribution shifts (e.g., hospital data from urban vs. rural settings) degrade performance catastrophically (Koh et al., 2021). Unlike biological systems, which evolved to prioritize causally salient features (e.g., predators, nutrients), ML lacks evolutionary pressure to distinguish signal from noise, rendering its causal cognition inherently shallow (Marcus and Davis, 2020).

3.2 Engineered causal architectures and the limits of extrinsic goals

Contemporary machine learning (ML) systems attempt to integrate causal reasoning through architectures that blend statistical learning with formal causal frameworks. Structural causal models (SCMs) represent one approach, applying Pearl's do-calculus (Pearl, 2019) to infer interventions from observational data. Tools like DoWhy (Sharma and Kiciman, 2020) and CausalNex based on Bayesian DAGs (Zheng et al., 2018) operationalize this by encoding causal graphs, yet they require human-specified variables and struggle with latent confounders—a critical limitation in real-world datasets (Schölkopf et al., 2021). For example, in healthcare, SCMs often fail to account for unmeasured socioeconomic factors that mediate treatment outcomes (Kaddour et al., 2022).

Causal reinforcement learning (CRL) extends this by training systems to learn intervention policies. DeepMind's Causal Meta-RL (Ke et al., 2022) demonstrates how systems can infer task structure through trial-and-error interventions, yet their objectives remain static (e.g., maximizing game scores). Unlike biological systems, which dynamically repurpose goals (e.g., switching from foraging to predator evasion), CRL systems lack mechanisms to reconfigure objectives in response to existential needs (Lake et al., 2017).

Neuro-symbolic hybrids merge neural networks with symbolic logic to enforce causal rules (Sheth et al., 2023). However, these rules are externally imposed rather than emergent from self-constitution, rendering them brittle under novel scenarios where "goal-directed commonsense is needed" (Garcez and Lamb, 2023). For instance, symbolic constraints in autonomous driving systems (e.g., "stop at red lights") fail to adapt when road conditions defy predefined norms (e.g., emergency vehicles). Table 1 summarizes the limitations of causality from our goal-oriented perspective.

These architectures reveal a fundamental limitation regarding their goals: MLS goals are extrinsic optimizations (e.g., loss minimization), and hence they decouple causal reasoning from

variables that underlie the actions their knowledge shall engage into Marcus and Davis (2020).

4 Key causal reasoning differences between autopoietic systems and machine learning systems

We now organize the differences between how Autopoietic Systems and MLS relate to causal reasoning into key dimensions that help to identify further venues for their scientific study.

TABLE 1 Training method for achieving goals and limitation examples.

Training paradigm	Causal limitation	Example	References
Supervised (accuracy)	Fails under distribution shift	Medical diagnosis	Geirhos et al., 2020
Reinforcement (reward)	No adaptive repurposing of objectives	Game score maximization	Ke et al., 2022
Neuro-symbolic (rules)	Brittle to novel scenarios needing commonsense	Autonomous driving	Garcez and Lamb, 2023

4.1 Operationalization of goals and causality

Causality in autopoietic systems is fundamentally recursive and self-referential, rooted in their operational closure. These systems' fundamental goal is to remain alive, i.e., maintain their identity through circular cause-effect chains that simultaneously produce and depend on their own boundaries (Maturana and Varela, 2012). For instance, a cell's metabolic network synthesizes the very components—enzymes, membranes, and organelles—that enable its continued existence. Here, causality is inseparable from the system's teleological imperative: self-preservation. Perturbations to autopoietic systems (e.g., nutrient deprivation) trigger adaptive responses aimed at restoring homeostasis, illustrating how causal reasoning is intrinsically directed toward sustaining systemic coherence (Luisi, 2003). The system's "goal" is not external but emergent of its self-reinforcing organizational structure.

In contrast, causality in machine learning (ML) systems is extrinsically defined by statistical correlations derived from training data. ML models, such as deep neural networks, infer patterns through gradient-driven optimization, with no inherent representation of counterfactuals or physical mechanisms (Pearl, 2019). For example, a convolutional neural network trained to classify images associates pixel configurations with labels (e.g., "cat" or "dog"), but these associations lack intrinsic grounding in the system's structure. The "goal" of an ML system—minimizing a loss function—is imposed externally by designers, reflecting no existential imperative. While autopoietic systems exhibit endogenous causality (causal processes emerge from self-maintenance), ML systems rely on exogenous causality, where causal attribution is bounded by the scope and biases of training datasets (Marcus, 2018) or rules imposed (Marcus and Davis, 2020). We summarize our discussion in Table 2.

4.2 Material embodiment and causal reasoning improvement

In autopoietic systems, causal reasoning improves through physical embodiment of increasingly complex goals. The material substrate available by its current operational organization, under the right mutations, directly enables the development of sophistication through what Heylighen (1995) calls meta-system transitions. These transitions enable new levels of control, i.e., the ability to handle perturbations in novel ways. This process is physically instantiated—cellular differentiation creates novel causal potentials through material reorganization that enables emergent functions. Major evolutionary transitions (Szathmáry, 2015) demonstrate how material reorganization drives causal advancement. When independent autopoietic units integrate into higher-order collectives, they physically restructure to enable a larger entity with new causal capabilities not only in relation to its environment but also with respect to its own components via top-down control (Rosas et al., 2020). This physical restructuring creates multi-level regulatory networks where causality operates across nested spatial and temporal scales—from molecular recognition (microseconds) to memory formation (decades).

This material integration enables the expansion of goal-directed structures across wider spatial domains and longer temporal horizons, termed the "care cone" (Witkowski et al., 2023). Long-term planning in mammals emerges not from computational scaling but from physical evolution of neural structures that materially integrate multiple internal states. In contrast, MLSs develop causal reasoning through pathways decoupled from their material embodiment. Their physical substrate remains static while abstract parameters adjust, creating a thermodynamic disconnect between energy expenditure and causal improvement. Neural networks expend identical energy regardless of whether they're refining causal models or reinforcing spurious correlations (Thompson et al., 2020).

This thermodynamic decoupling constrains how causal reasoning improves in MLSs. Without material reorganization that extends control across scales, causal advancement becomes purely instrumental—serving externally defined objectives without extending capacity for self-maintenance, or any other existential notion embedded in it. MLS causal improvements depend primarily on data quantity rather than material reorganization. While evidence suggests resemblances of major evolutionary transitions, by the identification of phase-transition-like improvements when crossing certain network size thresholds (Schölkopf et al., 2021), these remain qualitatively different from evolutionary transitions. ML systems expand within predefined architectural constraints without generating novel material configurations that enable fundamentally new forms of causal reasoning.

TABLE 2 Summary of operational differences between autopoietic and machine learning causal reasoning systems.

Dimension	Autopoietic systems	Machine learning systems
Causal mechanism	Recursive, based on self-referential loops at multiple hierarchical levels	Non-recursive, based on statistical correlations and gradient optimization
Teleological basis	Endogenous (self-maintenance)	Exogenous (loss minimization)
Boundary definition	Operational closure (self-produced membrane/organization)	Data distribution and algorithmic constraints

TABLE 3 Differences on how autopoietic and ML systems improve their causal cognition.

Aspect	Machine learning systems	Autopoietic systems
Mechanism of Improvement	Parameter optimization within fixed architecture	Material reorganization across multiple scales
Thermodynamic Relationship	Energy use unrelated to causal advancement	Energy expenditure aligned with control capabilities
Temporal integration	Reconstruction through increasingly smaller processing timescales	Nested timescales from molecular to evolutionary
Evolutionary potential	Bounded by architectural constraints	Open-ended through major transitions
Self-reorganization	No physical restructuring during learning	Continuous physical adaptation to maintain viability

The latter explains why biological causal reasoning exhibits open-ended improvement while artificial causality remains bounded by its instrumental nature and thermodynamic decoupling. We summarize our discussion in Table 3.

4.3 The attribution of causality

The epistemological foundations of causal attribution differ radically between autopoietic and machine learning systems, illuminating a philosophical dimension that combines their divergent material and operational principles.

4.3.1 Mechanistic causality in autopoietic systems

In autopoietic systems, causal attribution is mechanistic and grounded in material interactions. The experimental methods used to understand biological causality—such as knockout studies in yeast that reveal causal roles of genes in glycolysis (e.g., *PGI1* deletion halts glucose metabolism)—directly manipulate physical components to observe system-wide effects (Kitano, 2002), and can alter or be altered by features at other scales through mechanisms that might involve for example electrical, chemical and fluid-dynamics levels (Levin, 2019). The epistemology of causation is tied to physically observable and manipulable processes, such as enzyme-substrate binding observed via crystallography or electron microscopy.

This mechanistic approach to causality reveals the ontic nature of autopoietic causation—causal relationships are embedded in the system's physical structure and processes rather than being observer-dependent constructs. When biologists identify a gene as causal for a phenotype, they are identifying material pathways through which physical interactions propagate effects (Bechtel and Richardson, 2010). The multi-level organization of these systems means that causality operates simultaneously across molecular, cellular, and organismal scales, with bidirectional influences between levels.

4.3.2 Instrumental causality in machine learning systems

In contrast, causal attribution in ML systems is instrumental and *post hoc*. Researchers employ techniques such as ablation studies (removing a neural network layer) or Shapley values (Lundberg and Lee, 2017) to approximate feature importance, but these approaches provide statistical approximations rather than revealing physical causal mechanisms. When an ML researcher identifies a feature as "causal," they are making an epistemic claim about statistical relationships rather than identifying a physical pathway.

The epistemology of ML causality remains fundamentally statistical—a high Shapley value for a pixel in an image classification task doesn't imply a physical causal pathway but instead indicates a statistical correlation that the model has learned to exploit. This distinction becomes critical when deploying ML systems in real-world contexts where causal understanding (rather than correlation) is necessary for safe and effective operation.

5 Complementarity through evolutionary perspectives

Despite the fundamental differences between autopoietic systems and MLS outlined in previous sections, several promising pathways for complementarity are emerging. These approaches address the key gaps we've identified—particularly in goal formation, material embodiment, and causal understanding—offering potential bridges across this evolutionary divide.

5.1 Evolutionary algorithms and open-ended exploration

The application of evolutionary principles to machine learning design represents a significant step toward more biologically-inspired systems. Unlike traditional optimization approaches that focus on maximizing predefined metrics, evolutionary algorithms emphasize diversity, adaptation, and the emergence of novel solutions (Lehman and Stanley, 2011). Recent advances in quality diversity algorithms and Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) demonstrate how maintaining behavioral diversity rather than focusing solely on performance

leads to more robust and creative solutions that better resemble biological adaptability (Mouret and Clune, 2015; Cully et al., 2015).

Open-ended evolution research extends this approach by creating computational environments where continuous innovation emerges without predefined fitness functions (Taylor et al., 2016). These systems begin to address the goal-directedness gap identified in Section 4.1 by enabling the discovery of novel objectives rather than merely optimizing predefined ones. As Packard et al. (2019) note, truly open-ended artificial systems require mechanisms that support not just optimization but the continuous emergence of new ways to define and pursue goals—a property inherent to biological evolution (Boden, 1998).

However, these approaches still operate primarily at the algorithmic level, with limited impact on the material substrate gap highlighted in Section 4.2. The challenge remains integrating open-ended exploration with physical embodiment, a fundamental requisite for aitiopoietic cognitive systems.

5.2 Embodied cognition and material intelligence

A more direct approach to bridging the material gap involves developing artificial systems where materiality plays a constitutive role in cognition. Beyond conventional robotics, which often implements disembodied algorithms in physical shells, emerging research explores how material properties themselves can perform computational functions (Pfeifer et al., 2007; Paul, 2006).

Recent work in morphological computation demonstrates how physical dynamics can replace explicit computation, allowing systems to leverage material properties for intelligent behavior (Hauser et al., 2011; Müller and Hoffmann, 2017). Soft robotics and programmable materials enable adaptive behavior through their intrinsic material properties rather than through explicit programming (Laschi and Cianchetti, 2014; Rieffel et al., 2009). These approaches begin to address the thermodynamic decoupling identified in Section 4.2 by creating systems where physical structure directly participates in information processing.

The challenge remains developing materials that not only compute but also maintain and regenerate themselves—a defining aspect of autopoietic systems. However, the "aitiopoietic" integration of autopoietic structures into causal and other forms of reasoning is still in its infancy (McMillen and Levin, 2024).

5.3 Homeostatic regulation and predictive processing

The active inference framework (Friston, 2010; Friston et al., 2017) offers a computational approach that potentially bridges autopoietic self-maintenance and machine learning. By framing cognition as the minimization of surprise (or free energy) through continuous prediction and updating, this framework provides a computational account of how systems maintain homeostasis while adapting to environmental challenges.

Recent implementations in artificial systems demonstrate how predictive architectures can develop intrinsic goals related to maintaining viable states (Baltieri and Buckley, 2019). Particularly promising is research integrating homeostatic regulation directly into neural network architectures. Lechner et al. (2021) have demonstrated neural networks with homeostatic mechanisms that maintain internal stability while adapting to external challenges, exhibiting a primitive form of the intrinsic goal-directedness characteristic of autopoietic systems.

These approaches begin to address the causality gap identified in Section 4.3 by grounding causal understanding in the system's own viability conditions rather than in purely statistical correlations. However, they still operate primarily within the computational domain, with limited connection to physical self-maintenance. In this vein, conceptual advancement has been made by Kolchinsky and Wolpert (2018) by defining the concept of "semantic information" that refers to the Shannon information responsible for its viability, i.e., its self-production in autopoietic terms. However, this measure has not been yet properly linked to self-production but to proxy viability formulas that are informational as well.

5.4 Multi-scale integration and hierarchical agency

The hierarchical nature of autopoietic systems—with their nested levels of organization and regulation—finds a parallel in emerging approaches to multi-scale artificial intelligence. Recent advances in hierarchical reinforcement learning and multi-agent systems demonstrate how collective intelligence can emerge from interactions between simpler agents operating at different scales (Domingo-Fernández et al., 2022).

When designed with appropriate structural coupling between levels, these systems can develop emergent goals and coordination patterns reminiscent of biological collectives (Levin et al., 2023). This multi-scale approach potentially addresses the causal attribution gap identified in Section 4.3 by enabling both bottom-up and top-down causation across different levels of organization. This is consistent with evolutionary proposals that focus not only on individual but group selection and cooperation mechanisms, that have been proven useful in biology and culture (Wilson, 1975; Foster et al., 2017; Wilson et al., 2023), and being recently adopted as an alternative foundational paradigm in economics (Wilson and Snower, 2024).

6 Discussion

We examined the fundamental differences between autopoietic systems (biological organisms) and machine learning systems (MLSs), focusing on how their different natures affect causal cognition. We proposed that goals serve as a crucial pivot point for comparing these systems and explored potential convergence paths toward autopoietic systems that can integrate with MLS to embody "aitiopoietic cognition."

Our analysis proceeded by first establishing the apparent functional similarity but fundamental structural difference

TABLE 4 Comparative analysis of autopoietic systems and machine learning systems across operational, substrate, and epistemological dimensions.

Dimension	Autopoietic systems	Machine learning systems
Goal formation (operational)	Emergent from closure and self-maintenance imperatives	Externally imposed by designers
Causality implementation (operational)	Feedback-driven: within internal state via structural coupling	Statistical correlations and pattern recognition in data distributions
Physical embodiment of information (substrate)	Material substrate dynamically encode-decode information achieving computation through autopoietic processes	Static mapping between a specific physical part in charge of information states and other independent part in charge processing
Improvement process (substrate)	Physical reorganization across multiple scales enables enhanced control capabilities	Abstract parameter adjustment while maintaining fixed material architecture
Energy-cognition coupling (substrate)	Thermodynamic processes directly linked to cognitive enhancement and self-maintenance	Thermodynamic disconnect between energy expenditure and causal improvement
Causal knowledge source (epistemological)	Constitutional causality arising from material self-organization and boundary maintenance	Statistical inference from training data patterns and correlational structures
Causal attribution (epistemological)	Grounded in existential imperatives and viability conditions of the system itself	Based on computational optimization of externally defined loss functions
Learning foundation (epistemological)	Sense-making through autonomous environmental coupling and meaning generation	Information processing through supervised/unsupervised pattern extraction

between humans and MLSs cognition. We then examined autopoietic systems in depth, highlighting their self-organizing nature, emergent goals, and multi-level organization. Next, we analyzed how MLSs implement causal reasoning through statistical inference and optimization with externally imposed goals. This comparative framework allowed us to identify key differences in causal reasoning between these systems across three crucial dimensions: the operationalization of goals and causality, material embodiment and improvement mechanisms, and the epistemological foundations of causal attribution. Table 4 summarizes these fundamental operational, substrate, and epistemological differences, providing a concrete framework for understanding the evolutionary divide between biological and machine-learned causal systems.

The differences illustrated in Table 4 reveal profound implications for how we understand and develop artificial intelligence. Autopoietic systems exhibit recursive, self-referential causality rooted in their operational closure, with goals emergent from self-maintenance imperatives. Their causal reasoning improves through physical reorganization enabling multi-scale control, with material embodiment directly participating in cognition. In contrast, MLSs operate through extrinsic optimization and statistical correlations, with their physical substrate remaining static while abstract parameters adjust—creating a thermodynamic disconnect between energy expenditure and causal improvement.

TABLE 5 Convergence pathways between autopoietic and ML systems.

Convergence pathway	Addresses key gap	Current limitations	Exemplar research
Evolutionary algorithms	Goal formation	Limited material impact	MAP-Elites (Mouret and Clune, 2015)
Material intelligence	Physical embodiment	Lacks self-maintenance	Soft robotics (Laschi and Cianchetti, 2014)
Homeostatic regulation	Causal grounding	Primarily computational	Neural homeostasis (Lechner et al., 2021)
Multi-scale integration	Causal attribution	Limited embodiment	Multi-agent systems (Levin et al., 2023)

Our analysis suggest the need for integrating both cognitive architectures, including ML and autopoietic and perhaps others, with philosophical questions about causality and agency. Recent striking examples, reviewed in section "From Autopoiesis to Aitiopoiesis," demonstrate how these philosophical concepts find concrete expression in biological examples. Xenobots and organoids illustrate how collective behavior can emerge from cellular self-organization without external programming, suggesting that aitiopoietic properties can scale naturally from autopoietic foundations when appropriate organizational architectures emerge (Kriegman et al., 2020; Davies and Levin, 2023).

Therefore, the aitiopoietic cognition research program should proceed through four complementary tracks: (1) minimal aitiopoietic systems—engineering simple chemical/cellular systems that exhibit constitutional causality using synthetic biology techniques. For example, sensitivity and robustness of specific components might unveil causal-like behavior (Shinar et al., 2009); (2) measurement frameworks for characterizing aitiopoiesis—operationalizing goal-directedness by compatibilizing matter and information processing interplays. The latter requires an integration of notions from dynamical systems theory related to autopoiesis such as homeostatic recovery times, attractor basin and viability analysis, with information theoretical metrics explaining how the information of a system is processed in relation to its existence as a collective (Friston, 2010; Kolchinsky and Wolpert, 2018; Rosas et al., 2020); (3) scaling constitutional cognition—understanding how aitiopoietic properties emerge in a collective through multi-scale modeling approaches (Szathmáry, 2015); and (4) Hybrid Bio-Synthetic architectures—combining living materials with artificial substrates to achieve increasingly complex goal-directed behavior (Witkowski et al., 2023; Fields and Levin, 2022).

However, current convergence pathways face significant limitations: evolutionary algorithms encounter computational intractability in open-ended settings, material intelligence lacks genuine self-maintenance capabilities, homeostatic approaches remain primarily computational rather than constitutional,

and multi-scale integration struggles with embodiment challenges. These represent current research frontiers rather than insurmountable barriers, suggesting that breakthrough progress will require novel approaches that transcend these individual pathway limitations. These convergence challenges are summarized in Table 5.

7 Conclusion

The future of artificial intelligence may lie in systems that are neither fully machine nor fully living, but that bridge this evolutionary divide through novel forms of embodied, self-maintaining cognition. From here, we suggest that a research program aiming to reach "aitiopoietic cognition" should be developed. This ambitious research program convergence pathway involves the development of synthetic systems that exhibit genuine autopoietic properties—self-creation, self-maintenance, and self-boundary definition within their simulation domains. Research in synthetic biology and artificial life aims to create minimal chemical systems that display these properties (Luisi, 2003; Stano et al., 2023).

The realization of this research program will require studying metasystem transitions (Heylighen, 1995) and major evolutionary transitions (Szathmáry, 2015) in open-ended evolutionary settings. Metasystem transitions represent moments when systems develop new levels of control and coordination, fundamentally transforming their problem spaces (Witkowski et al., 2023) and enabling novel forms of causal reasoning. By creating artificial environments where such transitions can emerge naturally, we may navigate problem spaces with increasingly sophisticated "care-cones" (Witkowski et al., 2023)—expanding the spatial and temporal horizons over which they can exercise meaningful causal influence based on intrinsic rather than externally imposed goals (Deacon, 2011). This expansion of the care-cone would represent a crucial step toward artificial systems that can adapt to complex, open-ended environments through genuine understanding rather than mere statistical optimization (Seth, 2021). As Fields and Levin (2022) have argued, such systems would demonstrate competence across multiple domains through intrinsically motivated exploration and materially grounded causal reasoning (Walsh, 2015), potentially transforming our understanding of both biological and artificial intelligence.

This ambitious synthesis suggests a new understanding of cognition and agency that might bring us closer to resolving the mind-body problem, not through theoretical abstraction, but through the concrete development of systems that embody both material self-maintenance and sophisticated causal understanding.

The development of artificial entities with genuine intrinsic goals and materially grounded causal understanding raises fundamental ethical considerations that must be systematically addressed within this research program. To maintain analytical rigor within current scientific understanding, this discussion focuses on early-stage developmental issues rather than interesting but more speculative scenarios, which have been explored elsewhere (Kurzweil, 2005; Aerts and Arguëlles, 2022; Vidal, 2024). The proposed research program emphasizes minimal systems with constrained operational scope, necessitating the establishment of comprehensive safety protocols for self-modifying systems, ensuring beneficial alignment of early aitiopoietic prototypes that prioritize incremental progress with systematic monitoring of emergent capabilities.

A particularly significant ethical consideration involves understanding the fundamental material-informational dynamics underlying experiential states in general systems—an investigation that must remain central throughout the development of this research program (Witkowski et al., 2023). Given that aitiopoietic systems represent partially autonomous entities with potential experiential capacities, ethical protocols require continuous revision in accordance with evolving scientific understanding of agency, sentience, and consciousness, in artificial systems. This iterative ethical framework becomes essential precisely because the research program's object of study may possess forms of experience or proto-sentience that demand careful moral consideration as these systems develop greater autonomy and complexity.

This ambitious synthesis suggests a new understanding of cognition and agency that might bring us closer to resolving the mind-body problem, not through theoretical abstraction, but through the concrete development of systems that embody both material self-maintenance and sophisticated causal understanding. By developing artificial entities with truly intrinsic goals and materially grounded causal understanding, we may finally bridge the evolutionary divide between life and machine while maintaining commitment to careful, ethically-guided progress.

Author contributions

TV: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the UTEM grant LCLI23-01 entitled "Modelamiento de fenómenos interdisciplinarios complejos utilizando redes de reacciones" and the ANID grant no. 11241020 Fondecyt Iniciacion entitled "Modeling multidimensional disturbances and stability in ecology with reaction networks."

Acknowledgments

We would like to thank Selma Dundar-Coecke and Bob Coecke for the invitation to participate in this special issue, to Raphael Liogier, Olaf Witkowski and Stefan Leijnen for insightful discussions on the topic of this article, to the Systemic Modeling and Applications group at the Center Leo Apostel for internal discussion sessions, and to the reviewers for their dedicated reports.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. I used various LLMs to identify enhance my literature review, articulate subsections, and perform consistency checks regarding use of acronyms, grammar, and orthography.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aerts, D., and Arguëlles, J. A. (2022). Human perception as a phenomenon of quantization. *Entropy* 24:1207. doi: 10.3390/e24091207

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780195324259.001.0001

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv* [Preprint]. arXiv:1907.02893. doi: 10.48550/arXiv.1907.02893

Baltieri, M., and Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *Behav. Brain Sci.* 42:E218. doi: 10.1017/S0140525X19001353

Barandiaran, X., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386. doi: 10.1177/1059712309343819

Barandiaran, X., and Moreno, A. (2008). Adaptivity: from metabolism to behavior. *Adapt. Behav.* 16, 325–344. doi: 10.1177/1059712308093868

Barandiaran, X. E., and Egbert, M. D. (2014). Norm-establishing and norm-following in autonomous agency. *Artif. Life* 20, 5–28.

Bechtel, W., and Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research.* MIT press. doi: 10.7551/mitpress/8328.001.0001

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., et al. (2024). Managing extreme AI risks amid rapid progress. *Science* 384, 842–845. doi: 10.1126/science.adn0117

Boden, M. A. (1998). Creativity and artificial intelligence. *Artif. Intell.* 103, 347–356. doi: 10.1016/S0004-3702(98)00055-1

Boden, M. A. (1999). Is metabolism necessary? *Br. J. Philos. Sci.* 50, 231–248. doi: 10.1093/bjps/50.2.231

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Westport, CT: Praeger.

Clark, A. (2001). *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford: Oxford University Press.

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58, 7–19. doi: 10.1093/analys/58.1.7

Cully, A., Clune, J., Tarapore, D., and Mouret, J. B. (2015). Robots that can adapt like animals. *Nature* 521, 503–507. doi: 10.1038/nature14422

Davies, J., and Levin, M. (2023). Synthetic morphology with agential materials. *Nat. Rev. Bioeng.* 1, 46–59. doi: 10.1038/s44222-022-00001-9

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. New York, NY: W.W. Norton and Company.

Di Paolo, E., and Thompson, E. (2014). "The enactive approach," in *The Routledge Handbook of Embodied Cognition*, ed. L. Shapiro (London: Routledge), 68–78.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* 4, 429–452. doi: 10.1007/s11097-005-9002-y

Dittrich, P., and Speroni di Fenizio, P. (2007). Chemical organisation theory. *Bull. Math. Biol.* 69, 1199–1231. doi: 10.1007/s11538-006-9130-8

Domingo-Fernández, R., Alonso-Román, R., and Batalla-Fernández, L. A. (2022). Hierarchical reinforcement learning: a comprehensive survey. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3453160

Edelman, G. M., and Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proc. Nat. Acad. Sci.* 98:13763–13768. doi: 10.1073/pnas.231499798

Ellis, G. F. R. (2012). *How Can Physics Underlie the Mind?: Top-down Causation in the Human Context*. Cham: Springer Science and Business Media.

Fields, C., and Levin, M. (2022). Competency in navigating arbitrary spaces as an invariant for analyzing cognition in diverse embodiments. *Entropy* 24:819. doi: 10.3390/e24060819

Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Forgacs, G., and Newman, S. A. (2005). *Biological Physics of the Developing Embryo*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511755576

Foster, K. R., Schluter, J., Coyte, K. Z., and Rakoff-Nahoum, S. (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature* 548, 43–51. doi: 10.1038/nature23292

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Froese, T., and Ziemke, T. (2009). Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell.* 173, 466–500. doi: 10.1016/j.artint.2008.12.001

Gao, Y., Wang, L., and Wang, B. (2023). Customizing cellular signal processing by synthetic multi-level regulatory circuits. *Nat. Commun.* 14:8415. doi: 10.1038/s41467-023-44256-1

Garcez, A. D. A., and Lamb, L. C. (2023). Neurosymbolic AI: The 3 rd wave. *Artif. Intell. Rev.* 56, 12387–12406. doi: 10.1007/s10462-023-10448-w

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *Nat. Mach. Intell.* 2, 665–673. doi: 10.1038/s42256-020-00257-z

Gill, Z., Levin, M., and Veloz, T. (2025). Co-bootstrapping the origins of life: autonomy – pattern recognition cycles. *J. R. Soc. Interface.* (in review).

Goertzel, B. (2007). "Contemporary approaches to artificial general intelligence," in *Artificial General Intelligence, Vol. 2*, ed. C. Pennachin (New York, NY: Springer), 1–30. doi: 10.1007/978-3-540-68677-4

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridg, MA: MIT Press.

Hauser, H., Ijspeert, A. J., Füchslin, R. M., Pfeifer, R., and Maass, W. (2011). Towards a theoretical foundation for morphological computation with compliant bodies. *Biol. Cybern.* 105, 355–370. doi: 10.1007/s00422-012-0471-0

Heylighen, F. (1995). (Meta)systems as constraints on variation: a classification and natural history of metasystem transitions. *World Futures* 45, 59–85. doi: 10.1080/02604027.1995.9972554

Heylighen, F. (2023). The meaning and origin of goal-directedness: a dynamical systems perspective. *Biol. J. Linn. Soc.* 139, 370–387. doi: 10.1093/biolinnean/blac060

Huang, S. (2012). The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? *Bioessays* 34, 149–157. doi: 10.1002/bies.201100031

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/2528.001.0001

Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J., and Silva, R. (2022). Causal machine learning: a survey and open problems. *arXiv* [Preprint]. arXiv:2206.15475. doi: 10.48550/arXiv.2206.15475

Kauffman, S. A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford: Oxford University Press. doi: 10.1093/oso/9780195079517.001.0001

Ke, N. R., Bilaniuk, O., Goyal, A., Bauer, S., Racanière, S., Rezende, D. J., et al. (2022). Learning?to?induce?causal?structure.?*arXiv?preprint*?arXiv:2204.04875.

Kitano, H. (2002). Systems biology: a brief overview. *Science* 295, 1662–1664. doi: 10.1126/science.1069492

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., et al. (2021). "Wilds: a benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning* (PMLR), 5637–5664.

Kolchinsky, A., and Wolpert, D. H. (2018). Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* 8:20180041. doi: 10.1098/rsfs.2018.0041

Kriegman, S., Blackiston, D., Levin, M., and Bongard, J. (2020). A scalable pipeline for designing reconfigurable organisms. *Proc. Nat. Acad. Sci.* 117, 1853–1859. doi: 10.1073/pnas.1910837117

Kurzweil, R. (2005). "The singularity is near," in *Ethics and Emerging Technologies*, ed. R. L. Sandler (London: Palgrave Macmillan UK), 393–406. doi: 10.1057/9781137349088_26

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. *Behav. Brain Sci.* 40:E253. doi: 10.1017/S0140525X16001837

Laschi, C., and Cianchetti, M. (2014). Soft robotics: new perspectives for robot bodyware and control. *Front. Bioeng. Biotechnol.* 2:3. doi: 10.3389/fbioe.2014.00003

Lechner, M., Hasani, R., Amini, A., Henzinger, T. A., Rus, D., Grosu, R., et al. (2021). Neural circuit policies enabling auditable autonomy. *Nat. Mach. Intell.* 3, 791–799. doi: 10.1038/s42256-020-00237-3

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. doi: 10.1038/nature14539

Lehman, J., and Stanley, K. O. (2011). Abandoning objectives: evolution through the search for novelty alone. *Evol. Comput.* 19, 189–223. doi: 10.1162/EVCO_a_00025

Levin, M. (2019). The computational boundary of a "self": developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* 10:2688. doi: 10.3389/fpsyg.2019.02688

Levin, M., Bongard, J., and Lungarella, M. (2023). A framework for designing living machines. *Proc. Nat. Acad. Sci.* 120:e2221050120.

Lorenz, K. (1981). *The Foundations of Ethology*. Cham: Springer. doi: 10.1007/978-3-7091-3671-3

Luisi, P. L. (2003). Autopoiesis: a review and a reappraisal. *Naturwissenschaften* 90, 49–59. doi: 10.1007/s00114-002-0389-9

Lundberg, S. M., and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* 30.

Manzi, F., Peretti, G., Di Dio, C., Cangelosi, A., Itakura S., Kanda, T., et al. (2020). A robot is not worth another: exploring children's mental state attribution to different humanoid robots. *Front. Psychol.* 11:2011. doi: 10.3389/fpsyg.2020.02011

Marcus, G. (2018). Deep learning: a critical appraisal. *arXiv* [Preprint]. arXiv:1801.00631. doi: 10.48550/arXiv.1801.00631

Marcus, G., and Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. New York, NY: Pantheon.

Marcus, G., and Davis, E. (2020). The next decade in AI: four steps toward robust artificial intelligence. *arXiv* [Preprint]. arXiv:2002.06177. doi: 10.48550/arXiv.2002.06177

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., et al. (2025). *The AI Index 2025 Annual Report*. Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.

Maturana, H. R., and Varela, F. J. (2012). *Autopoiesis and Cognition: The Realization of the Living, Vol. 42*. Cham: Springer Science and Business Media.

McMillen, P., and Levin, M. (2024). Collective intelligence: a unifying concept for integrating biology across scales and substrates. *Commun. Biol.* 7:378. doi: 10.1038/s42003-024-06037-4

Moor, J. H. (1976). An analysis of the turing test. *Philos. Stud.* 30, 249–257. doi: 10.1007/BF00372497

Moreno, A., and Etxeberria, A. (2005). Agency in natural and artificial systems. *Artif. Life* 11, 161–175. doi: 10.1162/1064546053278919

Moreno, A., and Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Cham: Springer. doi: 10.1007/978-94-017-9837-2

Moro, C., Lin, S., Nejat, G., and Mihailidis, A. (2019). Social robots and seniors: a comparative study on the influence of dynamic social features on human–robot interaction. *Int. J. Soc. Robot.* 11, 5–24. doi: 10.1007/s12369-018-0488-1

Mouret, J. B., and Clune, J. (2015). Illuminating search spaces by mapping elites. *arXiv* [Preprint]. arXiv:1504.07614. doi: 10.48550/arXiv.1504.07614

Müller, V. C., and Hoffmann, M. (2017). What is morphological computation? On how the body contributes to cognition and control. *Artif. Life* 23, 1–24. doi: 10.1162/ARTL_a_00219

Nakagaki, T., Yamada, H., and Tóth, Á. (2000). Maze-solving by an amoeboid organism. *Nature* 407:470. doi: 10.1038/35035159

Newman, S. A., and Bhat, R. (2009). Dynamical patterning modules: a "pattern language" for development and evolution of multicellular form. *Int. J. Dev. Biol.* 53:693. doi: 10.1387/ijdb.072481sn

Packard, N., Bedau, M. A., Channon, A., Ikegami, T., Rasmussen, S., Stanley, K. O., et al. (2019). An overview of open-ended evolution: editorial introduction to the open-ended evolution II special issue. *Artif. Life* 25, 93–103. doi: 10.1162/artl_a_00291

Paul, C. (2006). Morphological computation: a basis for the analysis of morphology and control requirements. *Rob. Auton. Syst.* 54, 619–630. doi: 10.1016/j.robot.2006.03.003

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 54–60. doi: 10.1145/3241036

Pfeifer, R., and Lungarella, M., Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science* 318, 1088–1093. doi: 10.1126/science.1145803

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. New York, NY: William Morrow and Company. doi: 10.1037/e412952005-009

Priorelli, M., Stoianov, I. P., and Pezzulo, G. (2025). Embodied decisions as active inference. *PLoS Comput. Biol.* 21:e1013180.

Putnam, H. (1967). "Psychological predicates," in *Art, Mind, and Religion*, eds. W. H. Capitan, and D. D. Merrill (Pittsburgh, PA: University of Pittsburgh Press), 37–48. doi: 10.2307/jj.6380610.6

Rafelski, S. M., and Theriot, J. A. (2024). Establishing a conceptual framework for holistic cell states and state transitions. *Cell* 187, 2633–2651. doi: 10.1016/j.cell.2024.04.035

Rieffel, J. A., Valero-Cuevas, F. J., and Lipson, H. (2009). Morphological communication: exploiting coupled dynamics in a complex mechanical structure to achieve locomotion. *J. R. Soc. Interface* 11:20140520. doi: 10.1098/rsif.2009.0240

Rosas, F. E., Mediano, P. A., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., et al. (2020). Reconciling emergences: an information-theoretic approach to identify causal emergence in multivariate data. *PLoS Comput. Biol.* 16:e1008289. doi: 10.1371/journal.pcbi.1008289

Russell, S. J., and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach, 4th Edn*. London: Pearson.

Salthe, S. N. (1985). *Evolving Hierarchical Systems: Their Structure and Representation*. New York, NY: Columbia University Press. doi: 10.7312/salt91068

Schölkopf, B. (2022). "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl* (New?York, NY: Association?for?Computing?Machinery), 765–804. doi: 10.1145/3501714.3501755

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., et al. (2021). Toward causal representation learning. *Proc. IEEE* 109, 612–634. doi: 10.1109/JPROC.2021.3058954

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. New York, NY: Dutton/Penguin Random House.

Sharma, A., and Kiciman, E. (2020). DoWhy: an end-to-end library for causal inference. *arXiv* [Preprint]. arXiv:2011.04216. doi: 10.48550/arXiv.2011.04216

Sheth, A., Roy, K., and Gaur, M. (2023). Neurosymbolic artificial intelligence (why, what, and how). *IEEE Intell. Syst.* 38, 56–62. doi: 10.1109/MIS.2023.3268724

Shinar, G., Alon, U., and Feinberg, M. (2009). Sensitivity and robustness in chemical reaction networks. *SIAM J. Appl. Math.* 69, 977–998. doi: 10.1137/080719820

Smith, J. M., and Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Solé, R., Amor, D. R., Duran-Nebreda, S., Conde-Pueyo, N., Carbonell-Ballestero, M., Montañez, R., et al. (2016). Synthetic collective intelligence. *BioSystems* 148, 47–61. doi: 10.1016/j.biosystems.2016.01.002

Solé, R. Conde–Pueyo, N., Pla–Mauri, J., Garcia–Ojalvo, J., Montserrat, N., Levin, M. (2024). Open problems in synthetic multicellularity. *NPJ Syst. Biol. Appl.* 10:151. doi: 10.1038/s41540-024-00477-8

Stanley, K. O., and Lehman, J. (2015). *Why Greatness Cannot be Planned: The Myth of the Objective*. Cham: Springer. doi: 10.1007/978-3-319-15524-1

Stano, P., Nehaniv, C., Ikegami, T., Damiano, L., and Witkowski, O. (2023). Autopoiesis: Foundations of life, cognition, and emergence of self/other. *BioSystems* 232:105008. doi: 10.1016/j.biosystems.2023.105008

Szathmáry, E. (2015). Toward major evolutionary transitions theory 2.0. *Proc. Nat. Acad. Sci.* 112, 10104–10111. doi: 10.1073/pnas.1421398112

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv* [Preprint]. arXiv:1312.6199. doi: 10.48550/arXiv.1312.6199

Taylor, T., Bedau, M., Channon, A., Ackley, D., Banzhaf, W., Beslon, G., et al. (2016). Open-ended evolution: perspectives from the OEE workshop in York. *Artif. Life* 22, 408–423. doi: 10.1162/ARTL_a_00210

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

Thompson, N. C., Greenewald, K., Lee, K., and Manso, G. F. (2020). The computational limits of deep learning. *arXiv* [Preprint]. arXiv:2007.05558. doi: 10.48550/arXiv.2007.05558

Varela, F. J. (1979). *Principles of Biological Autonomy*. Amsterdam: North Holland.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/6730.001.0001

Veloz, T. (2021). Goals as emergent autopoietic processes. *Front. Bioeng. Biotechnol.* 9:720652. doi: 10.3389/fbioe.2021.720652

Veloz, T., and Razeto-Barry, P. (2017). Reaction networks as a language for systemic modeling: Fundamentals and examples. *Systems* 5:11. doi: 10.3390/systems5010011

Vidal, C. (2024). What is the noosphere? Planetary superorganism, major evolutionary transition and emergence. *Syst. Res. Behav. Sci.* 41, 614–622. doi: 10.1002/sres.2997

Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution* 50, 1008–1023. doi: 10.1111/j.1558-5646.1996.tb02342.x

Walsh, D. M. (2015). *Organisms, Agency, and Evolution*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9781316402719

Wang, P. (2019). On defining artificial intelligence. *J. Artif. Gen Intell.* 10, 1–37. doi: 10.2478/jagi-2019-0002

Weber, A., and Varela, F. J. (2002). Life after Kant: natural purposes and the autopoietic foundations of biological individuality. *Phenomenol. Cogn. Sci.* 1, 97–125. doi: 10.1023/A:1020368120174

West, S. A., Fisher, R. M., Gardner, A., and Kiers, E. T. (2015). Major evolutionary transitions in individuality. *Proc. Nat. Acad. Sci.* 112:10112–10119. doi: 10.1073/pnas.1421402112

West-Eberhard, M. J. (2003). *Developmental Plasticity and Evolution*. Oxford: Oxford University Press. doi: 10.1093/oso/9780195122343.001.0001

Wilson, D. S. (1975). A theory of group selection. *Proc. Nat. Acad. Sci.* 72, 143–146. doi: 10.1073/pnas.72.1.143

Wilson, D. S., Madhavan, G., Gelfand, M. J., Hayes, S. C., Atkins, P. W., Colwell, R. R., et al. (2023). Multilevel cultural evolution: from new theory to practical applications. *Proc. Nat. Acad. Sci.* 120:e2218222120. doi: 10.1073/pnas.2218222120

Wilson, D. S., and Snower, D. J. (2024). Rethinking the theoretical foundation of economics I: the multilevel paradigm. *Economics* 18:20220070. doi: 10.1515/econ-2022-0070

Witkowski, O., Doctor, T., Solomonova, E., Duane, B., and Levin, M. (2023). Toward an ethics of autopoietic technology: stress, care, and intelligence. *BioSystems* 231:104964. doi: 10.1016/j.biosystems.2023.104964

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: continuous optimization for structure learning. *Adv. Neural Inf. Process. Syst.* 31.

Ziemke, T., Bergfeldt, N., Buason, G., Susi, T., and Svensson, H. (2004). Evolving cognitive scaffolding and environment adaptation: a new research direction for evolutionary robotics. *Conn. Sci.* 16, 339–350. doi: 10.1080/09540090412331314821