Patterns of Purpose: From Dynamics to Goals

Roly Perera[1,3], Mads Hansen[2] and Soumya Banerjee[1]

[1]Department of Computer Science and Technology, University of Cambridge, UK
[2]Department of Philosophy, University of Bristol, UK
[3]School of Computer Science, University of Bristol, UK
roly.perera@cl.cam.ac.uk

Abstract

As artificial systems grow in autonomy, principled criteria are needed to determine whether a system genuinely pursues goals, and if so, which ones. We survey how different theoretical traditions — from control theory and dynamical systems to autopoiesis, reinforcement learning, and active inference — characterise goal-directedness, finding broad consensus that *persistence* and *plasticity* are key criteria, but disagreement on whether the underlying mechanisms should be understood in representational terms. We develop a pragmatic account that sidesteps this disagreement: we argue that the mechanistic structure underlying persistence and plasticity is always, in principle, empirically discoverable, and that this structure can pragmatically be treated as goal representation: if internal states correlate with preferred outcomes and the system adapts when the mapping between actions and outcomes changes, then it *represents* its goals in the sense that matters for prediction, explanation, and intervention. Such goal representations are typically distributed across multiple interacting components, as we illustrate with examples from evolved neural agents and neural cellular automata. Drawing on techniques from AI interpretability, we outline a two-stage experimental approach to reverse-engineering these distributed mechanisms, with ALife systems as tractable testbeds.

Keywords: agency, goal-directedness, persistence, plasticity, representation, distributed mechanism, interpretability

Submission type: **Full Paper**

1 Introduction

We often use teleological language — from Aristotle's *telos*, or goal — to describe systems that appear to act with purpose. Sometimes this is clearly metaphorical or anthropomorphic shorthand: a river might be said to be "seeking" the sea. But in many cases the language of goals seems to be indispensable; without it we lose predictive and explanatory traction on behaviour. Biological organisms are paradigmatically goal-directed, searching for food, avoiding injury, and regulating metabolic processes in ways that are naturally described as directed toward preferred outcomes.

In artificial systems, however, the notion of having a goal is more controversial. When does patterned, apparently future-directed behaviour constitute genuine pursuit of a goal, rather than mere predictable dynamics? Even in natural systems, it is not always obvious when complex dynamical behaviours are appropriately thought of as "goal-directed".

The problem is sharpened by the complementary risks of *over-attribution* and *under-attribution*. Humans are prone to over-attribute goal-directedness: in the classic experiments of Heider and Simmel (1944), simple moving geometric shapes were spontaneously described as pursuing, fleeing, or deceiving one another, a reflection of cognitive biases tuned to detect intentional action. Large language models exacerbate this tendency by producing fluent natural-language reports about goals without necessarily meeting any substantive criteria for goal-directedness (Quattrociocchi et al., 2025).

Under-attribution also carries risk: treating genuinely goal-directed systems as passive tools risks failing to anticipate misaligned behaviour (Shapira et al., 2026; Amodei et al., 2016; Russell, 2022). Principled criteria for evaluating the goal-directedness of artificial agents, determining what their goals might be, and where appropriate, intervening to shape them, are therefore urgently needed (Marchal et al., 2026).

These concerns are most pressing in AI, for obvious reasons. In the longer term, they are likely to become pressing in artificial life too, as ALife systems grow in autonomy and the boundary between the two fields continues to blur. But ALife also has a more immediate contribution to make: its bottom-up approach, in which lifelike dynamics are constructed from simple components under precise experimental control, offers a testing ground for developing the principled criteria needed more broadly.

Developing such criteria requires a clearer theoretical account of what goal-directedness is. But the ubiquity of goal-directed behaviour has historically presented a challenge, namely how to reconcile goals with a fundamentally non-teleological physical universe, first raised in Schrödinger's (1944) *What is Life?* Contemporary perspectives increasingly frame the issue in terms of open-system thermodynamics and self-organisation. These "physics-friendly" perspectives characterise goal-directedness as a way of interpreting and explaining the dynamics of systems that actively maintain themselves in far-from-equilibrium states.

While these perspectives provide a useful naturalistic framing, many important questions remain. If a system exhibits genuine goal-directedness, are those goals necessarily *represented* in (or by) the system? What might this mean operationally, in terms of reverse-engineering them? What even is *genuine* goal-directedness, and can it be meaningfully distinguished from *apparent* goal-directedness? Goal-directedness is thus both an *explanans* — a widely used and apparently indispensable explanatory construct — and an *explanandum* — a substantive natural phenomenon that itself demands mechanistic explanation and theoretical clarification.

In this paper we argue that the field of artificial life has much to contribute to this debate, and is likely to gain new theoretical traction in return. We first draw on the philosophical and scientific literature to set out a taxonomy of goal-directed systems, and then develop a pragmatic account of goal representation. We argue that the mechanistic processes underlying the key attributes of goal-directedness, namely *persistence* and *plasticity*, are always, in principle, empirically discoverable, and propose operational criteria and experimental paradigms for doing so. We draw on work in artificial life, where lifelike dynamics have been studied computationally for decades, but also take inspiration from mechanistic interpretability techniques in AI, which, under similar pragmatic assumptions to ours, have yielded considerable success at reverse-engineering representations.

2 Characterising Goal-Directedness

The study of goal-directed behaviour appears across many disciplines: from philosophy, evolutionary biology, control theory, cybernetics and dynamical systems theory, to more recent fields such as reinforcement learning and active inference. Despite substantial differences in formalisms and explanatory aims, these traditions converge on two core features characteristic of goal-directedness (Nagel, 1979):

1. **Persistence:** the tendency for a system following a particular behavioural trajectory to return to that trajectory following perturbations that cause it to depart.

2. **Plasticity:** the tendency for a system to reach a particular behavioural trajectory from a variety of different starting points or environmental conditions.

Both criteria are inherently *modal*: they describe not just what the system actually does, but what it *would* do across a range of counterfactual circumstances (Nagel, 1979). To say that a cat is *pursuing* a mouse is to say that the cat would have taken a different route had the mouse gone a different way; more generally, in a goal-directed system the means counterfactually depend on the goal (Walsh, 2012). It is this modal character — robustness of outcome across variation in circumstances and means — that distinguishes genuinely teleological explanation from mere description of actual behaviour.

Some authors use these terms in a rather weak sense, to mean any system that can be seen as directed towards an end (Babcock and McShea, 2021; Babcock, 2023). A hot object cooling to the ambient temperature, or a ball settling in a bowl, might in this weak sense be called "plastic" (reaching the same end from different starting conditions). A ball returning to the bottom after being nudged might similarly be called "persistent". But such *convergent* systems are not actively regulating their behaviour: the ball has no alternative means of getting there, and does no work to do so. Convergent systems are directed towards an end only in the sense that their trajectories are constrained by an energy landscape; they lack the adaptive flexibility that we take to be characteristic of genuine persistence and plasticity.

More typically, persistence and plasticity are used in a strictly richer sense, to refer to *non-convergent* systems: systems that do active work to maintain or bring about an outcome under varying circumstances[1]. This is the usage we assume here. However, there is a strict notion of persistence and plasticity for non-convergent systems which ties goal-directedness specifically to *self-maintenance* (Deacon, 2011). Indeed, self-maintaining systems are the canonical cases of goal-directedness; such systems remain far from thermodynamic equilibrium by actively doing work to counteract entropy increase. Living cells and organisms are the paradigmatic examples, maintaining themselves in a "viability region" of state space through self-repair, homeostatic regulation and other corrective mechanisms. Although these systems are still ultimately dissipative, they are non-convergent. Such dynamics have also been studied in artificial systems: from Maturana and Varela's (1980) original simulations of autopoiesis to progressively richer models exhibiting growth, self-repair, and reproduction (McMullin, 2004). Fontana and Buss's (1994) artificial chemistry showed how self-maintaining molecular structures can arise spontaneously in certain computational substrates.

Importantly, not every system in a far-from-equilibrium state qualifies as self-maintaining. Some systems exhibit characteristics that at face value look like self-maintenance. A whirlpool re-forms after small disturbances, but it does not act to preserve the temperature gradient on which its existence depends; it is convergent on a non-equilibrium attractor, not self-maintaining. More generally, there is a broad class of self-organising systems, including tropical cyclones, candle flames, and methuselahs in Conway's Life (Gardner, 1970), which spontaneously form organised structures far from equilibrium, but do not act to preserve the conditions that sustain them, depending instead on external driving.

Persistence and plasticity come in degrees: a bacterium tumbling up a chemical gradient has a narrow behavioural repertoire, while humans draw on cognitive, social, and technological strategies to pursue goals under diverse and novel

[1] Such systems have also been called *targeted* (García-Valdecasas, 2025; Levesley et al., 2025)

circumstances. Regardless, persistence and plasticity together are widely accepted as sufficient for goal-directedness: the various traditions broadly agree that a system exhibiting both is goal-directed.

We take a somewhat more permissive view. Although self-maintaining systems are plausibly the evolutionary origin of all goal-directed behaviour, end-directed behaviours that exhibit adaptive flexibility (plasticity in the choice of means) without being directed towards self-maintenance are also, on our account, genuinely goal-directed. A cell in a multicellular organism, although itself self-maintaining, also exhibits behaviours, such as apoptosis or immune signalling, that serve the organism rather than the cell itself. A person planning a card game may exhibit considerable plasticity, but the connection to their survival is at best indirect; an LLM embedded in an agentic framework may pursue goals with adaptive flexibility without being self-maintaining at all. In such cases the goals the system serves may ultimately be those of a larger system of which it is a component, rather than "its own" in any strong sense.

One point of difference among these traditions is whether such non-convergent behaviour necessarily involves *goal representation*: an internal encoding of preferred states that guides behaviour. Some accounts (such as control theory and active inference) take representation to be integral to the dynamics of persistence and plasticity; others (such as field theory) regard it as unnecessary or even misleading. A further question, then, is whether the mechanisms underlying persistence and plasticity can or should be understood in representational terms.

Table 1 summarises how these ideas have been articulated across traditions. In **control theory and cybernetics**, goal-directedness is identified with the use of *feedback* to maintain or restore a target condition under environmental perturbation (Wiener, 1948; Ashby, 1956). Formally, a system is specified by a target state (or *set point*) r, plus an output $y(t)$ and error term $e(t) = r - y(t)$ representing the deviation from r at time t; goal-directedness consists in using negative feedback to minimise $e(t)$. The Good Regulator theorem (Conant and Ashby, 1970) implies that any such regulator must model the system being regulated, regardless of how that model is encoded; in this sense goal-directed control necessarily involves representation.

In **dynamical systems theory**, goal-directedness arises when the state space is organised around a stable regime (attractor or viability set), together with processes that can be interpreted as *regulatory*, i.e. as preserving the conditions under which the system continues to occupy those states (Prigogine and Stengers, 1985; Kauffman, 1993). Goals correspond to the regions toward which trajectories converge; unlike simple passive convergence, such as a marble settling in a bowl, goal-directedness involves a coupling between system and environment such that perturbations trigger compensatory dynamics that maintain convergence. This account is largely agnostic about representation: one could argue that the mechanisms steering the system back towards an attractor under perturbation need not involve an explicit model of the preferred state, and may simply be inherent in the dynamics themselves.

Evolutionary biology typically takes an *etiological* (causal-historical) stance towards goal-directed behaviour, explaining it in terms of evolutionary history. Behaviours are goal-directed inasmuch as they were selected for producing certain effects; those historically selected effects then determine present standards of success and failure, giving such views a normative dimension (Millikan, 1984, 1989). Tying goal-directedness to evolutionary history can be explanatorily useful; sometimes surprising behaviours can be understood as adaptations to past environments, for example. But etiological accounts of goal-directedness may be less applicable to synthetic life forms or artificial systems that lack evolutionary histories of their own (Levin, 2022). A related but distinct position is Mayr's *teleonomic* view (Mayr, 1961, 1974): apparently goal-directed processes are those which execute a "program": a mechanistic process, such as embryonic development, that has been shaped by evolution to produce a certain outcome. This is an instrumentalist take that treats goal-directedness as a useful gloss on underlying mechanism, and proponents would most likely resist the idea that goals are genuinely represented by the system.

In the theory of **autopoiesis**, organisms are autonomous systems whose actions are structured by the need to preserve their own organisation; behaviour is goal-directed insofar as it contributes to viability and continued self-production (Maturana and Varela, 1980). **Enactivism** extends this to a characterisation of cognition as "embodied sense-making", whereby the organism enacts or embodies a specific coupling to the world through adaptive self-regulation (Varela et al., 1992; Thompson and Varela, 2001). On this view, goals are not internal representations that precede action but are implicit in the organism's ongoing interactions with its environment. One might question whether this leaves sufficient room for temporally extended planning or explicit anticipation of future states (Wheeler, 2017); arguably enactivists need not deny goal representation per se but only the idea that goals are inherently internal, brain-bound states.

In **reinforcement learning** (RL), goal-directed behaviour is the process of selecting actions to maximise expected cumulative reward (Sutton et al., 1999). Agents learn policies that map states to actions by estimating *value functions* encoding this long-term expected reward for different courses of action; goals enter the system via the *reward* function, which assigns higher value to some outcomes over others. In model-free variants, the learned policy implements behaviour directly; in model-based variants, behaviour is mediated through a *transition model* for more deliberative, planning-based control. Model-based RL is overtly representational, but even model-free agents represent goals implicitly: the

Framework	Persistence	Plasticity	Representation
Control theory / cybernetics	Negative feedback sustains the system near the set point despite ongoing perturbation	The system can reach the set point from a range of initial conditions	Yes: Good Regulator theorem implies regulator must model the regulated system
Dynamical systems theory	Compensatory dynamics return the system to its attractor after perturbation	Trajectories converge on the attractor from a variety of initial states	Agnostic: regulatory mechanisms may be inherent in the dynamics rather than encoded in an internal model
Biology (etiological)	Homeostatic behaviours, self-repair, and other mechanisms sustain the organism's viability	Phenotypic plasticity: organism reaches viability from range of environmental and genetic conditions	Not required: goal-directedness grounded in selection history, not internal states
Teleonomy (Mayr)	Evolved "programs" reliably execute to completion despite perturbation	Programs can produce the same outcome from varied environmental conditions	Resisted: goal-directedness is a gloss on mechanism, not genuine representation
Autopoiesis / enactivism	Self-production maintains the system's organisation despite perturbation	The organism can achieve viability from a range of initial and environmental conditions	Resisted: goals implicit in organism–environment coupling
Reinforcement learning	Learned policies sustain reward-maximising behaviour despite environmental changes	Agents can learn effective policies from varied initial states and reward structures	Yes: explicit in model-based RL (transition models); implicit in model-free RL (value functions)
Teleodynamics (Deacon)	Reciprocal constraints sustain the system away from equilibrium	Composite system can arise from varied initial configurations of its self-organising components	Yes: A system represents a future state by being disposed toward realising such a state
Field theory (Babcock & McShea)	External field sustains system's trajectory despite perturbation	Field guides system towards same outcome from varied starting points	No: source of goal-directedness is always external to the entity
Active inference (Friston)	Both action and belief revision minimise gap between system and preferred states	Flexible capacity to reach viability enabled by organism's ability to construct its own niche	Yes: goals are priors with explicit representational content

Table 1: How different frameworks interpret persistence, plasticity, and goal representation

learned value function encodes which states are preferable, even though the agent has no explicit model of what it is trying to achieve. In both cases the reward function is usually externally given, and the question of how to specify reward functions that lead to desirable behaviours remains an open problem (Amodei et al., 2016; Russell, 2022).

The **teleodynamics** framework (Deacon, 2011) characterises goal-directed behaviour as a higher-order reciprocal relationship between self-organising processes, in which each process creates the boundary conditions, or *constraints*, that prevent the other from dissipating, sustaining the composite system away from thermodynamic equilibrium. Constraints are central to this account: what a teleodynamic system is disposed towards is not a state, but the preservation of a *constraint structure*: a set of reciprocal boundary conditions that channel work in a way that sustains the whole (García-Valdecasas and Deacon, 2024). The system represents its end, on this view, by being so organised that its processes tend to reproduce the very constraints upon which they depend, not by anything that functions as an internal model of a goal state.

The **field theory** (Babcock and McShea, 2021; Babcock, 2023) is an explicitly externalist account: goal-directed activity arises when an entity is embedded in a spatially extended, physically real *field* that persistently and plastically guides its trajectory. The term "field" is used broadly, encompassing fields in the familiar physical sense (gravitational, electromagnetic) but also (for example) a morphogenetic gradient that guides embryonic development, or an ecological selection pressure landscape that directs adaptation. Because the source of goal-directedness is understood as external, no internal representation of goal states is required on this account.

Finally, the **active inference** framework (Friston et al., 2006; Friston and Stephan, 2007; Friston, 2009) attempts to subsume several of the above traditions. Goal-directed behaviour is the selection of actions in order to minimise expected (variational) free energy within a generative model of the environment. Agents maintain probabilistic "beliefs" (probability densities) about states of the external world and the sensory consequences of actions; action and perception (belief revision) then provide two complementary ways to reduce prediction error relative to those beliefs. "Goals" in this framework are priors over expected sensory outcomes that bias action selection toward certain states (Friston et al., 2017), and thus are very explicitly internal states with representational content: the preferred outcomes.

3 Representing Goals

As the preceding survey shows, the various traditions broadly agree that persistence and plasticity are the hallmarks of

goal-directedness, but disagree on the role of representation, with positions ranging from the instrumentalism of Mayr (1961) to more realist views. One might worry that further progress requires first settling the question of whether persistent and plastic systems "really" represent their goals or are merely usefully described as such. We argue that no such prior settlement is needed. Instead, we can focus on what is empirically discoverable: the mechanisms underlying the specific behaviour we are inclined to see as goal-directed, and the interventions that modify them. This *operationalises* the notion of goal representation, allowing progress regardless of ontological commitments: if a system exhibits adaptive flexibility, responding to novel conditions in ways that advance a preferred outcome, then positing that it represents its goals becomes a useful explanatory strategy, enabling prediction and intervention.

The search for goal representations then plays two distinct roles. For systems already taken to be uncontroversially goal-directed, it helps us determine *what* the goal is, and potentially how to manipulate it: for example, bioelectric interventions on planaria reveal that the organism represents a target morphology which can be experimentally rewritten (Levin, 2022). For systems whose dynamics are uncertain — including artificial systems but potentially borderline naturally occurring systems too — the search for goal representations serves as a classification procedure: if a system can be shown to represent preferred outcomes, that is strong evidence of adaptive flexibility, and hence of goal-directed behaviour.

This approach is compatible with Dennett's (1991) idea of a "real pattern": if a system's goal-directed behaviour is robust enough that one could make money betting on one's ability to predict and control it, then — by a "no-miracles"-style argument (Putnam, 1975) — there must be systematic and discoverable mechanisms underlying that behaviour, regardless of ontological commitments. A complementary argument comes from ALife: Rocha (1998) argues that self-organisation alone, bound to a fixed attractor landscape, cannot account for the open-ended adaptability of evolved systems. Some form of representational structure (what he calls "selected self-organisation") is needed to explain adaptation to genuinely novel circumstances, and embodiment does not remove this need (Rocha, 1998, p. 352).

We therefore propose the following working definition:

> A **goal representation** is an internal or embodied structure or process encoding one or more preferred states of the world, together with mechanisms that select actions expected to increase the likelihood of those states.

The proviso "internal or embodied" captures our desire to remain neutral about whether goals are represented in a specialised internal structure such as a brain (Kristan, 2016) or in a more system-wide fashion. What one is looking for, concretely, is evidence that internal states or processes correlate with preferred external states or outcomes — a *correlational* criterion that identifies *candidate* goal representations — and then evidence that those candidates are causally implicated in the system's competences, through *counterfactual sensitivity*: that the system adapts its policies when the mapping between actions and outcomes is altered, rather than merely continuing with a fixed behavioural routine. Counterfactual sensitivity is a way of exploring the modal structure of goal-directed behaviour noted in § 2: by intervening on the system or its environment, we probe what it *would* do under altered circumstances, in the spirit of interventionist accounts of causation (Woodward, 2003).

3.1 Distributed representations

However, reverse-engineering goal representations is likely to be a significant challenge, because the mechanisms orchestrating goal-directed behaviour are rarely localised in a single component. In most natural and engineered systems, these mechanisms are distributed across multiple interacting subsystems, so one cannot simply open up the system and find salient structures representing goals. This distributed organisation has also been called *heterarchical* (McCulloch, 1945). This is a familiar challenge in AI, where knowledge is spread across millions of parameters in a neural network, none of which is individually meaningful; in artificial systems generally, this motivates interpretability work that spans multiple components rather than focusing on any single one.

Work in ALife provides a concrete illustration. Beer's (2003) analysis of evolved CTRNN agents performing categorical perception, catching circular objects and avoiding diamond-shaped ones, found no evidence of circle detectors, corner detectors, or other internal representations in the traditional sense. The agent's competence was achieved entirely through the coupled dynamics of brain, body, and environment. Beer's analysis thus challenges a traditional notion of representation, in which readily isolable internal states stand in for external categories.

However, the same analysis revealed that certain perceived features, notably object width and the presence or absence of corners, were major determining factors in the agent's discrimination, meaning its responses were selectively sensitive to those features. On our pragmatic account, the mechanisms which implement this selectivity count as a distributed form of goal representation: the coupled dynamics that underpin the agent's competences must have some mechanistic structure discoverable by empirical investigation, regardless of one's willingness to label this structure "representational". The case therefore illustrates not the absence of goal representation, but the importance of looking to more computationally realistic notions of representation. Indeed, subsequent information-theoretic analysis of the same class of agents (Williams and Beer, 2010) confirms that task-relevant information is distributed across the agent's internal states, and

that mutual information between internal and environmental variables tracks the agent's discriminative competence: precisely the kind of mechanistic structure our account predicts should be discoverable.

That goal selection and control are distributed in this way, with no natural central locus for decision-making, has been observed independently across cybernetics (Ashby, 1956), cognitive architecture (Minsky, 1986), behaviour-based robotics (Brooks, 1986), and the philosophy of neuroscience (Bechtel, 2021). Indeed, the difficulty of assigning functional responsibility to unique components is the norm wherever multiple subsystems interact; Table 2 collects some familiar examples.

System type	How goal-directedness is distributed
Genome and epigenome	Behavioural dispositions emerge from gene–environment interaction rather than being transparently encoded in DNA
Neural circuits	Goal-directed behaviour emerges from circuit-level dynamics rather than being localised in individual neurons
Morphology and movement	Responsibility for locomotion distributed across brain, skeleton, and local ganglia
Culture and environment	Tools and social institutions scaffold goal-directed behaviour across individuals
Social groups	Coordination emerges from interaction among agents, often without central control

Table 2: Distributed goal-directed organisation across scales

Distributedness also raises questions about where to draw the boundaries of the "system" whose goal-directedness is at issue. As with representations, system boundaries are also partly a function of what we want to explain: different explanatory agendas may carve out different systems. For self-maintaining systems there may be more principled ways of individuating; *constraint closure* (Montévil and Mossio, 2015) draws the boundary around those component processes that mutually depend on and generate the constraints sustaining one another, forming a self-maintaining loop across multiple timescales, similar to Deacon's (2011) teleodynamics. Whatever participates in the closure counts as part of the system.

But many systems of interest resist such clean demarcation, and here the perspectival character of the boundary becomes evident. Symbotic organisms (Dupré and O'Malley, 2009) may have goals that are not directed towards their own maintenance but to the maintenance of the other, so that self-maintenance is then a property of the composite system rather than of either component. In extended versions of Lenia (Chan, 2019), distinct self-organised structures can coexist and interact symbiotically, maintaining their integrity through cooperation rather than in isolation, raising similar questions about how to individuate the relevant system. A robot that pursues goals but does not engage in self-repair or recharging may exhibit rich goal-directed behaviours that serve the self-maintenance of a larger system, namely its human master; whether one treats the robot or the human–robot composite as "the system" depends on the explanatory question at hand.

The same holds for LLMs taking on more autonomous roles: In isolation an LLM merely draws from probability distributions over token sequences, but one embedded in an agentic framework that interprets outputs as commands, executes code, and triggers actions in the world forms part of a composite system capable of pursuing goals (Wang et al., 2024). As with genes, which become causally potent only when embedded in an interpretive system of transcription machinery and regulatory networks (Griffiths and Stotz, 2013), the question of where to locate goal-directedness depends on where one draws the system boundary.

3.2 Interpretability and pluralism

The question of whether, or how, a system represents a particular goal is not only complicated by distributedness but is in a deeper sense a matter of interpretation. If we want to characterise representations by what they *do*, namely by how they contribute to the functional economy of the system at large, then the answer will depend on the explanatory or interventional agenda we bring to the inquiry.

This is not merely the observation that different analytical frameworks (say, active inference and dynamical systems theory) can be legitimately applied to the same system (Beer and Williams, 2015). Rather, different *explananda* — different behavioural phenomena or competences of interest — induce genuinely different accounts of what the system is doing representationally. A system that catches circles and avoids diamonds may, when probed for a different competence, reveal an entirely different representational organisation. This is not to deny the existence of objective facts altogether regarding such questions (as an anti-realist might), but to say something rather less deflationary: simply that such facts are always relativised to a particular perspective. This is akin to Bongard and Levin's (2023) *polycomputing*, the idea that the same substrate can be interpreted as doing many different things at the same time. Zhang et al. (2025) show that even very simple, well-understood components, even traditional sorting algorithms like bubblesort, can exhibit unexpected competencies when embedded into a larger system, supporting entirely new readings of what the component is, contextually, doing.

Recent work in AI interpretability illustrates a complementary aspect of this pluralism. Using empirical interpretability methods, Kim et al. (2026) claim that the enhanced capabilities of "chain of thought"-style reasoning in LLMs can be explained by understanding the system as implicitly simulating complex multi-agent discourse. A simpler example would be recent work on helical representations of numbers in LLMs (Kantamneni and Tegmark, 2025), where it is shown

that LLMs' non-trivial arithmetic competencies can be understood as operations on these implicit representations. In both cases, what makes the interpretation significant is not that it is the unique "correct" reading of the system, but that it is empirically productive: it enables prediction, explanation, or intervention. For goal representation specifically, the interpretive question is sharpened in systems like LLMs embedded in agentic frameworks, where goal-relevant information may be distributed across model weights, context window, and scaffolding, so that reverse-engineering it requires studying the system's dynamics in its operating environment rather than internal structure in isolation.

Our suggestion is that goal-directedness should be approached from a similar pluralistic perspective, so that the question of whether a machine exhibits goal-directed behaviour, regardless of what its designers intended, boils down to a purely empirical question that can only be settled by considering the capabilities of the machine in the context of some larger system in which it is embedded. Different interventional agendas will carve up goal structures differently; the perspectival stance is therefore not merely philosophical but methodologically necessary.

3.3 Discovering goal representations

Given this pragmatic and pluralistic framing, namely that goal representations are in principle discoverable, typically distributed, and always relative to an explanatory agenda, what concrete methods can we bring to bear? Scientific experimentation on a system can be understood as a controlled form of the interpretive activity just described: we place the system into specific contexts and observe how it acts, creating opportunities to interpret internal mechanisms or structures as representational.

The two kinds of evidence identified previously — correlational and interventional — suggest a corresponding two-stage experimental programme. The first stage identifies *candidate* representations: internal states or processes that correlate with the system's goals. This can proceed either backwards from observed behaviour, inferring a representation of preferred states that would best explain the observed pattern of action selection, or forwards from internal structure, examining internal states directly for correlates of task-relevant behaviours. The second stage confirms that the candidates so identified are causally implicated in the system's competences, through intervention. Mechanistic interpretability research on AI systems has developed concrete methods for both stages (Bereska and Gavves, 2024).

Working backwards from behaviour, inverse reinforcement learning (Ziebart et al., 2008; Finn et al., 2016; Hadfield-Menell et al., 2016) has shown that reward functions can be reliably inferred from observed action sequences alone. Working forwards, observational methods such as *probing* — training classifiers on a model's internal activations to detect encoded information — and *sparse autoencoders* — decomposing neural activations into interpretable features (Bricken et al., 2023) — identify internal states that correlate with task-relevant variables, providing candidate representations whose causal role can then be tested interventionally. A central challenge is that individual neurons are typically *polysemantic*, encoding multiple unrelated concepts: a concrete instance of the representational pluralism discussed in § 3.2, and a challenge shared with neuroscience, where individual biological neurons similarly encode multiple stimulus features. Sparse autoencoders address this by decomposing activations into a larger set of monosemantic features, though the decomposition is not unique: different configurations yield different feature sets, reinforcing the point that representational readings depend on the analytical tools employed and what one is trying to explain.

The interventional side of this programme has developed rapidly in AI. *Activation patching* replaces specific internal activations and observes effects on behaviour, testing which components are causally necessary or sufficient for a given competence (Conmy et al., 2023). *Automated circuit discovery* extends this to identify the minimal computational subgraph — the "circuit" — responsible for a specific behaviour, by iteratively pruning connections whose removal does not affect performance. More recently, *circuit tracing* (Ameisen et al., 2025) builds interpretable replacements for language models, decomposing polysemantic neurons into monosemantic features and tracing feature-to-feature causal contributions via attribution graphs.

For LLMs embedded in agentic frameworks, a particularly revealing interventional test is to place the system in environments with competing objectives: a system that weighs these against one another, and revises its priorities when expected payoffs change, demonstrates the kind of counterfactual sensitivity that, on our account, constitutes strong evidence for goal representation. All these methods arguably operate under the same pragmatic assumption that we propose: what matters is not whether internal structures or mechanisms are best labelled "representations", but whether they are causally implicated in the system's competences.

ALife systems offer particularly tractable testbeds for this two-stage programme, since they allow precise experimental control over both the system and its environment. Several of the ALife studies discussed in § 3.1 already instantiate this two-stage logic. Williams and Beer's (2010) information-theoretic analysis of evolved CTRNN agents is essentially observational: it measures mutual information between internal and environmental variables, identifying candidate representations by establishing that the agent's dynamics encode discriminative information about its world. Beer's (2003) stimulus-perturbation experiments are the interventional counterpart, demonstrating that specific dynamical couplings are causally responsible for discrimination. Crutchfield and Mitchell (1995) provide the closest ALife analogue to circuit discovery: by identifying homogeneous regions in

the spatiotemporal behaviour of evolved cellular automata and subtracting them, they reveal propagating boundary structures ("particles") whose collisions implement an embedded logic, isolating the important computational structure from the background dynamics.

To illustrate how this programme might be applied going forward, consider neural cellular automata (NCAs; Mordvintsev et al., 2020). An NCA learns local update rules that enable a grid of cells to grow a target morphology from a single seed, and crucially, to regenerate that morphology after damage. This self-repair capacity is a clear instance of persistence and plasticity: the system reliably achieves the same outcome across a range of perturbations, with no explicit goal encoded anywhere in the update rule. The two-stage programme would proceed as follows. First, probe the learned update rule to identify which features of the local neighbourhood state correlate with damage-relevant variables: for example direction and magnitude of deviation from the target pattern, or cell-state gradients that signal the boundary of a lesion. This is directly analogous to probing an LLM's activations for task-relevant features. Second, intervene: patch the candidate damage-direction signal to a fixed value and observe whether repair becomes disoriented, or ablate a candidate feature and test whether growth still converges to the target. If ablating a feature disrupts repair but not growth, that feature is causally implicated specifically in the persistence capacity. Whatever mechanistic structure this analysis isolates *is* the goal representation, on the pragmatic account developed above, because it is causally responsible for the system's counterfactual robustness.

Other recent ALife systems present similar opportunities. Hamon et al. (2025) use automated search methods to discover cellular automata rules that give rise to localised structures exhibiting primitive sensorimotor agency: the resulting structures move, react coherently to obstacles, and maintain integrity under perturbation, without any of these capabilities being explicitly programmed, again raising the question of the mechanistic structure underlying these competences. In all these cases, these experimental methods offer a principled way to explore the mechanisms underlying persistence and plasticity.

The analogy between AI interpretability and ALife can only be pushed so far: AI benefits from architectures with discrete, parameterised components that can be individually targeted for intervention, whereas ALife systems, like cellular automata and CTRNNs, are often closer to substrates than architectures. Despite these structural differences, both domains potentially support the same methodological core: identify a competence, probe for internal correlates, intervene to establish causal role, and isolate the minimal structure responsible. This convergence supports our broader claim that the mechanistic structure underlying goal-directed behaviour is, in principle, always empirically accessible.

4 Conclusion

As artificial systems become increasingly autonomous, the question of whether they genuinely pursue goals — and if so, which ones — is no longer merely theoretical. Beyond the obvious implications for ethics and governance, it also raises questions about mechanism: how does goal-directedness emerge where it was not explicitly engineered, and what methods can reveal the implementation techniques?

The pragmatic account we have developed is intended to make these questions tractable. Rather than asking whether a system "really" represents its goals, we focus on what is empirically discoverable: if a system exhibits persistence and plasticity, then the mechanistic structure underlying these capacities can in principle be found, and this structure can pragmatically be treated as goal representation. Because goal-directed behaviour is inherently modal — a claim about what the system would do across counterfactual circumstances — the natural way to investigate it is through intervention, and the experimental paradigms we propose are designed with this in mind. The chief difficulty is that the relevant mechanisms are typically distributed across multiple interacting components, so that no single part of the system can be identified as "the goal". Work in ALife, where persistence and plasticity can be studied under precise experimental control, offers a way to develop and refine these methods before applying them to natural and artificial systems where the stakes are highest.

We have not yet applied these methods to a concrete system; the NCA case study of § 3.3 sketches what such an application would look like. Carrying it through is the most immediate next step, with NCAs and similar ALife systems offering plausible starting points. A longer-term challenge is to bridge the gap between ALife testbeds and large-scale AI systems, where there is much work on reverse-engineering representations but relatively little on studying the representation of emergent goals.

We also set aside the question of *self-aware* systems. People — historically our starting point for thinking about goal-directedness — not only pursue goals but also reflect upon and reason about them (Dennett, 2012, 2004). By a conservative generalisation of the Good Regulator theorem, any self-regulating system must employ a self-model; if that self-model becomes rich enough for the system to interpret *itself* as a goal-directed agent, we might see the beginnings of self-awareness, including the ability to explicitly share, hide and revise goals themselves, not just the means by which they are pursued. There is already evidence of limited introspection in LLMs (Lindsey, 2026), and of minimal self-models developing when such systems are coupled with robotic interfaces (Yoshida et al., 2024). The relationship between self-representation and open-endedness in ALife systems thus also represents a natural extension of the experimental programme we have outlined.

References

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. (2025). Circuit Tracing: Revealing Computational Graphs in Language Models.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. *ArXiv*, abs/1606.06565.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. J. Wiley, New York.

Babcock, G. (2023). Teleology and function in non-living nature. *Synthese*, 201(4):112.

Babcock, G. and McShea, D. W. (2021). An externalist teleology. *Synthese*, 199(3-4):8755–8780.

Bechtel, W. (2021). Grounding cognition: Heterarchical control mechanisms in biological cognition. *Philosophical Psychology*, 34(4):482–510.

Beer, R. D. (2003). The Dynamics of Active Categorical Perception in an Evolved Model Agent. *Adaptive Behavior*, 11(4):209–243.

Beer, R. D. and Williams, P. L. (2015). Information Processing and Dynamics in Minimally Cognitive Agents. *Cognitive Science*, 39(1):1–38.

Bereska, L. and Gavves, E. (2024). Mechanistic Interpretability for AI Safety — A Review. *Transactions on Machine Learning Research*.

Bongard, J. and Levin, M. (2023). There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines. *Biomimetics (Basel, Switzerland)*, 8(1).

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N. L., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models with Dictionary Learning.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, 2(1):14–23.

Chan, B. W.-C. (2019). Lenia: Biology of Artificial Life. *Complex Systems*, 28(3):251–286.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, volume 36.

Crutchfield, J. P. and Mitchell, M. (1995). The Evolution of Emergent Computation. *Proceedings of the National Academy of Sciences*, 92(23):10742–10746.

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. A Touchstone Book. WW Norton & Company.

Dennett, D. C. (1991). Real Patterns. *The Journal of Philosophy*, 88(1):27–51.

Dennett, D. C. (2004). *Freedom Evolves*. Penguin Publishing Group.

Dennett, D. C. (2012). The free floating rationales of evolution. *Rivista di filosofia, Rivista quadrimestrale*, (2/2012):185–200.

Dupré, J. and O'Malley, M. A. (2009). Varieties of Living Things: Life at the Intersection of Lineage and Metabolism. *Philosophy & Theory in Biology*, 1:1–25.

Finn, C., Levine, S., and Abbeel, P. (2016). Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 49–58, New York, New York, USA. PMLR.

Friston, K. (2009). The Free-Energy Principle: A Rough Guide to the Brain? *Trends in Cognitive Sciences*, 13(7):293–301.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1):70–87.

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience & Biobehavioral Reviews*, 77:388–402.

Friston, K. J. and Stephan, K. E. (2007). Free-energy and the brain. *Synthese*, 159(3):417–458.

García-Valdecasas, M. (2025). Living systems are targeted: A challenge to the teleology of field theory. *Biology & Philosophy*, 40(2):8.

García-Valdecasas, M. and Deacon, T. W. (2024). Origins of Biological Teleology: How Constraints Represent Ends. *Synthese. An International Journal for Epistemology, Methodology and Philosophy of Science*, 204:1–28.

Griffiths, P. and Stotz, K. (2013). *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative Inverse Reinforcement Learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Hamon, G., Etcheverry, M., Chan, B. W.-C., Moulin-Frier, C., and Oudeyer, P.-Y. (2025). Discovering Sensorimotor Agency in Cellular Automata Using Diversity Search. *Science Advances*, 11(44).

Heider, F. and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243–259.

Kantamneni, S. and Tegmark, M. (2025). Language Models Use Trigonometry to Do Addition.

Kauffman, S. A. (1993). *The Origins of Order: Self Organization and Selection in Evolution*. Oxford University Press.

Kim, J., Lai, S., Scherrer, N., y Arcas, B. A., and Evans, J. (2026). Reasoning Models Generate Societies of Thought.

Kristan, W. B. J. (2016). Early evolution of neurons. *Current Biology*, 26(20):R949–R954.

Levesley, N., McShea, D. W., and Babcock, G. (2025). Evolving systems and directionality. *Interface Focus*, 15(5):20250018.

Levin, M. (2022). Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds. *Frontiers in Systems Neuroscience*, Volume 16 - 2022.

Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models.

Marchal, N., Chan, S., Franklin, M., Revel, M., Keeling, G., Fischli, R., Chandra, B., and Gabriel, I. (2026). Architecting Trust in Artificial Epistemic Agents.

Maturana, H. and Varela, F. J. (1980). *Autopoiesis and Cognition : The Realization of the Living*. Boston Studies in the Philosophy of Science ; v. 42. D. Reidel Pub. Co., Dordrecht, Holland ; Boston.

Mayr, E. (1961). Cause and Effect in Biology. *Science*, 134(3489):1501–1506.

Mayr, E. (1974). Teleological and Teleonomic, a New Analysis. In Cohen, R. S. and Wartofsky, M. W., editors, *Methodological and Historical Essays in the Natural and Social Sciences*, pages 91–117. Springer Netherlands, Dordrecht.

McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *Bulletin of Mathematical Biophysics*, 7(2):89–93.

Millikan, R. (1989). In Defense of Proper Functions. *Philosophy of Science*, 56(June):288–302.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Minsky, M. (1986). *The Society of Mind*. Simon & Schuster, New York.

Montévil, M. and Mossio, M. (2015). Biological Organisation as Closure of Constraints. *Journal of Theoretical Biology*, 372:179–191.

Mordvintsev, A., Randazzo, E., Niklasson, E., and Levin, M. (2020). Growing Neural Cellular Automata. *Distill*, 5(2).

Nagel, E. (1979). *Teleology Revisited and Other Essays in the Philosophy and History of Science*. Columbia University Press, New York Chichester, West Sussex.

Prigogine, I. and Stengers, I. (1985). *Order out of Chaos : Man's New Dialogue with Nature*. Fontana Paperbacks, London, flamingo edition. edition.

Putnam, H. (1975). *Mathematics, Matter and Method: Philosophical Papers*, volume 1. Cambridge University Press, Cambridge.

Quattrociocchi, W., Capraro, V., and Perc, M. (2025). Epistemological Fault Lines Between Human and Artificial Intelligence.

Rocha, L. M. (1998). Selected Self-Organization and the Semiotics of Evolutionary Systems. In *Evolutionary Systems: Biological and Epistemological Perspectives on Selection and Self-Organization*, pages 341–358. Springer.

Russell, S. (2022). Artificial Intelligence and the Problem of Control. pages 19–24.

Schrödinger, E. (1944). *What Is Life? & Mind and Matter: The Physical Aspect of the Living Cell*. Cambridge University Press.

Shapira, N., Wendler, C., Yen, A., Sarti, G., et al. (2026). Agents of Chaos.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Thompson, E. and Varela, F. J. (2001). Radical Embodiment: Neural Dynamics and Consciousness. *Trends in Cognitive Sciences*, 5(10):418–425.

Varela, F., Rosch, E., and Thompson, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press. MIT Press.

Walsh, D. (2012). Mechanism and Purpose: A Case for Natural Teleology. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43:173–181.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J.-R. (2024). A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6).

Wheeler, M. (2017). The Revolution Will Not Be Optimised: Radical Enactivism, Extended Functionalism and the Extensive Mind. *Topoi Orient – Occident*, 36(3):457–472.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, MA, 2 edition.

Williams, P. L. and Beer, R. D. (2010). Information Dynamics of Evolved Agents. In *From Animals to Animats 11: Proceedings of the Eleventh International Conference on Simulation of Adaptive Behavior*. Springer.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yoshida, T., Baba, S., Masumori, A., and Ikegami, T. (2024). *Minimal Self in Humanoid Robot "Alter3" Driven by Large Language Model*, volume ALIFE 2024: Proceedings of the 2024 Artificial Life Conference of *ALIFE 2022: The 2022 Conference on Artificial Life*.

Zhang, T., Goldstein, A., and Levin, M. (2025). Classical sorting algorithms as a model of morphogenesis: Self-sorting arrays reveal unexpected competencies in a minimal model of basal intelligence. *Adaptive Behavior*, 33(1):25–54.

Ziebart, B. D., Maas, A., Bagnell, J. A., and Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, pages 1433–1438, Chicago, Illinois. AAAI Press.