Cohering Reinforcement Learning

Anna Harutyunyan, Will Dabney, Doina Precup

Keywords: Foundations, philosophy, reward-is-not-enough, coherence

Summary

This is a philosophico-technical journey through the foundations of RL, in the pursuit of a new paradigm. Starting from the bedrock Markov assumption, we shake loose our certainty in the primacy of reward maximization when it does not hold, and trace a path through the structure of behaviour that remains uncaptured. We propose an alternative principle, and a simple but foundational technical modification that can express it as well as provide a fresh lens onto classical problems.

Contribution(s)

1. We elaborate on the implications of the Markov assumption, and provide a philosophical and conceptual argument for why reward maximization alone is insufficient for intelligent behaviour, when it does not hold.
   Context: This is a refinement of the "Reward is enough" hypothesis (Silver et al., 2021) for the reality of non-Markov worlds.
2. We introduce **coherence** – the drive towards a constitutive, resonant coupling between the agent and the environment – as a necessary principle in non-Markov worlds, complementary to reward maximization.
   Context: Our notion of coherence is drawn from and resonant with a number of disciplines, from continental philosophy to physics to cognitive science.
3. We propose a simple yet foundational shift – of the domain of operation from **real to complex numbers**, and conceptually show that it naturally offers mathematics to express coherence.
   Context: This work sketches the basis of this shift, leaving the complete technical and theoretical development for subsequent work.

Cohering Reinforcement Learning

Anna Harutyunyan, Will Dabney, Doina Precup
harutyunyan@google.com

Google DeepMind

Abstract

The Markov assumption is woven through the entire foundation of reinforcement learning, allowing it to be cast as a purely instrumentalist framework of maximizing reward. How does our framework change if we take state construction – necessary for non-Markov spaces – as a non-trivial part of intelligent behaviour? We make a case that another, non-instrumental principle becomes necessary. We propose *coherence* – a resonant coupling between the agent and the environment. In order to operationalize it meaningfully, we propose to expand our domain *from real to complex numbers*. The dynamics of the imaginary dimension provide natural mathematics to express the ideas we propose. We consider the seminal extended mind thesis as a useful case study. Bridging multiple disciplines, from continental philosophy to physics, this work sketches a promising direction for the next generation of agents that are concerned with more than their gain, and opens new pathways of thinking about the reinforcement learning problem.

1 Introduction

At the center of this work is the question of state construction. The standard reinforcement learning (RL) framework assumes the agent to be a passenger in an existing Markov world (Sutton & Barto, 2018). Deep RL turns it around – asking the agent to (ideally) learn a Markov state internally. In both cases, there is a strict separation between the agent and the environment – all the necessary information is contained either in the environment, or in the agent. We instead invite the focus onto the agent-environment *relationship*. The agent does not navigate a fixed world, or strive for an abstract ideal, but constantly co-creates with the environment.

Organisms construct niches, people externalize cognition. The agent is not a detached entity, and its ability to synergize with the environment is actually crucial for its success. The difference with the classical perspective is akin to that between "hard" and "soft" robotics (e.g., Pfeifer & Bongard, 2006) – achieving a pre-defined trajectory solely via internal motors vs. cooperating with gravity to the same end. In the latter case, the agent and the environment form "a unit" that moves together; in the former the agent operates entirely separately. Less of a laboratory, and more of a dance, this process of forming and moving as a unit is what we are interested in here, and we view it through the lens of *state*.

In RL, we usually think of state construction as representation learning: the agent works to learn a "good" representation. Behaviour is aimed at solving the task, or improving this internal state. We draw attention to an alternative view – behaviour that is *constitutive* of and aimed at producing a "good" agent-environment configuration. In line with this, Kirsh (1995) offers the example of "rearranging the workspace" before work, terming that type of behaviour as *"spatial"*. Contrasting it with *temporal* planning – finding an optimal sequence of actions "in time" – Kirsh argues that spatial action has been overlooked but is inalienable to natural intelligence, which is necessarily situated "in space".[1] Spatial behaviour is not merely transient until the agent learns what it needs to internally, but is part and parcel of its strategy of being in the world, constitutive to it. It is also subjective: the workspace is organized relative to a particular agent. Philosophically, this view is *phenomenological*: instead of implicitly aiming toward an ideal "out there," in third person, the agent inhabits its first-person perspective and acts to support its current reality.

We wish to articulate the principle of action responsible for inducing spatial behavior. Firstly, we posit that it is distinct from reward maximization. One might object – in line with debates around the distinction between epistemic and pragmatic action (Kirsh & Maglio, 1994) – that a difference in behavior does not necessarily imply a distinct driving principle. We argue in the affirmative, from a computational and philosophical perspective.

The principle we put forward is **coherence** – a drive for a resonant coupling between the agent and the environment, the "inside" and the "outside". We provide a view that the Markov assumption is in fact a special case of this, one where the agent and environment are structurally independent, and the coupling is achieved trivially. A useful related notion is that of *epistemic tension, or torque* – the cognitive resistance arising when integrating new concepts into an ontology (e.g., Sher, 2016). The reader may be experiencing it throughout this paper, for example. The drive towards coherence is the reduction of this tension. Unlike the related notion of free energy minimization (Friston, 2010), which may famously lead agents to a dark room with "no alarms and no surprises", our notion of coherence is not driven to *avoid* incoherence, but to resolve *as much of it as possible*, which implicitly includes experiencing it (so as to then resolve).

We consider the **relationship between features and weights** as the structural hinge for these ideas, and propose a radical yet simple shift – to extend the domain of operation **from real to complex numbers**. This allows us to consider relational dynamics; the imaginary dimension becomes a remarkably well-suited store for the "spatial" epistemic dimension, and the mechanics of the bridge between the imaginary and the real provide a way to express coherence mathematically.

As a concrete thread in our investigation, we consider the classical extended mind (EM) thesis from cognitive science (Clark & Chalmers, 1998). Living precisely in the relationship of the internal and the external, it challenges their boundary, drawing a provocative equivalence between using external devices and internal cognitive tools, like a notebook and biological memory. In the language we develop here – the behavior leading to incorporating the external in this way, is a special class of spatial action. We conceptually validate our notion of coherence, by phrasing EM in its terms and showing that our formalism provides an intuitive illustration for the thesis.

This work presents a step away from the classical objectivist view into a subjective and relational model of agent-environment interaction. The proper accounting of the epistemic and the inclusion of the imaginary provides a bridge towards computational notions of meaning for RL, while remaining comfortably tethered to familiar technical concepts. On the other hand, it links the RL model to classical physics and control theory, opening up new ways of understanding the RL problem. We believe this work to be the beginning of a promising path for developing foundations for the next generation of agents that are truly equipped to relate with the human and more-than-human world.

On this journey, we make meaningful connections between several disciplines, like continental philosophy, physics, cognitive science, and control theory – all through the vehicle of the generality of the RL problem, which has always been a nexus of interdisciplinary insight.

2 No Markov, no max

We begin at the foundational assumption of the classical RL framework: the Markov state. The future only depends on the present; the current state is all you need to know an action's possible effects. There is no need for memory, and the agent need never worry about state construction.[2]

[1] We interpret "space" here as the general inhabiting of an observation space, rather than strictly physical embodiment.

[2] Environmental Markovness can be obscured by the agent's computational or perceptual limitations, which we treat simply as a form of non-Markovness.

The entire remaining project is about identifying the correct course of action that navigates this state landscape to desired ends.

As Kirsh (1995) argues, this kind of *temporal* planning (finding an optimal sequence of actions "in time") has overshadowed an equally vital dimension of natural intelligence: *spatial* action. For situated beings, effective action requires more than traversing a world; it requires *organizing the workspace* to construct a better state. This behavior is neither optional nor transient, but an inalienable aspect of being in the world.

The Markov assumption implies a deep form of independence: not only conditional independence *in time* (between the present state and historical trajectory), but an ontological independence *in space* – a clean boundary between state and action, agent and environment. This is precisely what allows for the exclusive focus on "temporal" action. When the assumption does not hold, state and action become entangled. Actions no longer just traverse a fixed state landscape, but actively shape it and affect the agent's coupling with the environment, they acquire a spatial dimension. If the state was Markov, this would not be necessary; the workspace would already be organized.

2.1 Against instrumentalism

The key question is: are spatial actions, taken to "organize the workspace", instrumental? In this paper, we take the position that the underlying principle is distinct from instrumentalism, and "reward is not quite enough" (Silver et al., 2021).

To do this, we draw a distinction between *descriptive* and *normative* explanation: the question if spatial action ultimately leads the agent to maximize reward, and the question whether maximizing reward is the driving principle of spatial action, are in general separate. So while the result may (and should) be reward maximizing, we claim that the spatial principle *allows* for it, rather than emerges from it. It is constitutive, rather than instrumental in nature – aimed at state construction and compensating for the absence of a Markov state.

We ground this point in the practical reality of RL: **the credit assignment problem of non-Markov spaces**. Actions that organize the workspace or write notes do not result in external reward directly, but affect the agent's coupling with the environment that can lead to better reward later. Learning this can be particularly difficult though – the classical Bellman equation interprets non-Markovness as noise, and the utility of structural preparation is heavily obscured. If the agent is lucky it may learn workspace-organizing policies eventually, through random exploration and statistical averaging, but doing so robustly and with some sophistication is simply not afforded by the standard model. The practice of deep RL algorithms that ubiquitously includes auxiliary representation-shaping (state-constructing) objectives supports this.

Thus we have that (1) in the absence of a Markov state, the consequent entanglement of state and action gives rise to spatial behavior, and (2) spatial behaviour is not well-explained by reward maximization; leading us to claim that **in the absence of a Markov state, reward maximization alone is insufficient for learning intelligent behavior, and another, complementary principle is needed.**

2.2 Partial observability

Now consider the usual way of modeling non-Markovness – *partial observability*. This assumes that there is a perfect Markov state "out there", the agent just doesn't know it yet. The assumption is epistemic (a question of knowledge). The agent is concerned with gathering sufficient statistics to approximate this ideal unknown state.

Although there is comfort in the existence proof – a state space composed of the history of all observations from the beginning of time, is Markov, and hence can potentially be approximated – this model maintains the complete detachment between the agent and the environment. The environment is describable as Markov in some ideal, agent-independent way that the agent maintains beliefs about.

The phenomenological assumption is in contrast agent-centric and ontological (a question of existence) – there is *no* agent-independent reality that is relevant to the agent's experience. The agent is in a process of active co-creation with the environment. The resulting state is agent-specific and subjective. This does not require us to assume Markovness, or measure against it.

Partial observability methods tend to focus *retroactively* on state estimation and reconstructing histories. The state construction approach we advocate for focuses *proactively* on policy enactment – using spatial action to shape a meaningful landscape.

2.3 Philosophy

The history of clear and convenient boundaries traces most prominently to Descartes – between subject and object, mind and body, mental *"res cogitans"* and material *"res extensa"* (Descartes, 1641). Temporal planning, the problem of optimal action, the agent itself, is analogous to the mental; while state construction, spatial action, and the environment – to the material. We invite the reader to challenge and blur such boundaries for themselves.

Heidegger (1927) in "Being and Time" distinguished between states that are "ready-to-hand" *(Zuhanden)*, seamless and actionable for the agent, and "present-at-hand" *(Vorhanden)*, existing outside of the agent, detached and independent. Ready-to-hand states center interaction and are entwined with the agent's subjectivity. Markov states in contrast are always present-at-hand. Challenging the Cartesian divide, Heidegger prioritised readiness-to-hand, and argued that our practical engagement with the world as a *network of meaningful equipment* is more fundamental than theoretical, detached observation, and that the objective view (presence-at-hand) is merely a derivative abstraction.

Key to this work is the phenomenological perspective. *Phenomenology is an attempt to describe the basic structures of human experience and understanding from a first person point of view, in contrast to the reflective, third person perspective that tends to dominate scientific knowledge (Merleau-Ponty, 1962).* The agent is inside the experience of non-Markovness, rather than observing it from the outside.

3 Extended Mind

We examine a concrete class of behavior that fits the "spatial" description. Consider the seminal Extended Mind thesis (Clark & Chalmers, 1998) from cognitive science. The premise is that despite the temptation to consider cognition as contained within the physical boundaries of brain and skull, it in fact routinely *extends* into the world. We think not just in our heads, but "with" the world. The running example from the original paper involves two museum-goers Inga and Otto, making their way to MoMA. Inga remembers the address, while Otto has Alzheimer's disease, and writes everything in a notebook that he carries with him all the time. The authors claim that if the notebook is always accessible to Otto, there is no difference between Inga's way of storing information in her internal memory, and Otto in his external notebook.

The immediate insight for RL is that *learning via internal representation (e.g., memory) and learning via spatial policy (e.g., manipulating an external notebook) can serve the same ends.* Importantly, extended actions can be preferred not only if internal representation is impossible, but because the spatial route is simply much "easier" – long arithmetic is a good example – even if potentially suboptimal in terms of instrumental value (requiring more steps, for example).

A related point was brought up by Kirsh & Maglio (1994) (in the precursor to "The Intelligent Use of Space"). The authors consider the contrast of *pragmatic* and *epistemic* action, and discuss the example of Tetris. They highlight that even master players resort to rapid, redundant rotation of incoming pieces, rather than performing it in their head, which is technically possible. Mildly suboptimal, this strategy clearly prioritises a good agent-environment coupling.

Thus, we want to understand robust mechanisms of learning extended and spatial policies, not only out of philosophical obligation, but practical gain. Nature is well-versed in this, and boundary-blurring behaviors make up an important pillar of natural intelligence. Spiders build their nets to externalize sensation, ants leave pheromone trails as collective memory. Most prolific of all at extension are humans (Fernando et al., 2024), powered by their capacity for symbols and the symmetry of *engrams and exograms* (e.g., Sutton, 2010), language itself being an example of this. Arguably, **extension is a mechanism by which biological agents negotiate their physical and computational boundaries, and surpass their specific constraints by collaborating with their environment and continuously restructuring it towards fruitful integration.** It is fundamentally a *scaling mechanism*: less about making the agent bigger, and more about the agent-environment system becoming "better coupled".

On the quest of understanding the mechanics by which it comes about, we will now enrich the picture and go beyond the conventional interpretation of EM. We first review the classical objections to the thesis, then proceed to describe its "three waves".

Objections. Like any self-respecting theory, EM has had its share of opposition, perhaps most staunchly by Adams & Aizawa (2001; 2010, and related). The first objection is their rejection of what they call the *coupling-constitution fallacy*, whose meaning is conveniently self-evident: coupling $\neq$ constitution, the notebook remains hopelessly outside. A second key objection, also by Adams (2010) is what they refer to as a *lack of the "mark of the cognitive"* in extended objects, notebook containing merely "dead symbols". These objections have in turn sparked further responses from the community (e.g. Clark, 2010; Kagan & Lassiter, 2013; Piredda, 2017), producing a lively debate. We will touch upon them briefly in our developments.

3.1 Riding the waves

We now recount what is known as the *three waves* of the EM thesis.

First wave. The original thesis focused on *equivalence*: extended behavior is functionally equivalent to an internal capability. This is sometimes referred to as the "parity principle" (Clark & Chalmers, 1998). While intuitive, it relies on a loose common-sense functionalist epistemology of similarity. It is common to interpret this wave as *computational offloading*: the agent uses the world for cheaper computation.

Second wave. The years following the original thesis sparked much critical debate about more precise criteria qualifying extension (Sutton, 2010; Menary, 2010; Andrada, 2021). Instead of functional equivalence, this second wave emphasizes *complementarity*: it's not that the external process need implement the same function as its (fictional) internal equivalent, but that the processes need to *couple* productively (Sutton, 2010). This naturally includes equivalence but refines and expands it usefully, targeting attention on the coupling itself.

Third wave. Functionalism can still creep in into the second wave view, if we see the world as a passive hard drive available to couple with. The third wave departs from this by considering the *nature* of the coupling, and placing it as *sense-making* (Gallagher, 2013; Kirchhoff & Kiverstein, 2018; Prosen, 2022). This is the enactive perspective (Thompson & Stapleton, 2009): the agent continuously collaborates with the world to "make sense" of it, the "sense" is formed at and distributed across the boundary. The agent's subjectivity itself extends into the world. In this view, the "mark of the cognitive" from the second objection, is no longer limited to the agent's interior. The symbols in the notebook are no longer "dead".

3.2 A max in need is a max indeed

What little representation EM-like ideas have found in RL firmly sits in the first-wave category. The agent offloads computation into the environment (Risko & Gilbert, 2016), or implements missing internal functions due to a limited representation class (Deverett et al., 2019). These views, falling under the umbrella of bounded rationality, have their useful place, but come with their own burdens, and do not offer a significant philosophical departure from the norm.

In addition to requiring the agent to *perceive all relevant computation costs*, computation-offloading can be interpreted as *second-order* instrumentalism, and arguably replaces one difficult optimization problem with another. Indeed, it can literally be analyzed by constructing a meta-MDP (e.g., Griffiths et al., 2015), or implementing a value-based decision-making scheme (Gilbert, 2024). Hence, the philosophical considerations articulated in Section 2 remain applicable, only amplified by a danger of infinite regress.

We wish to follow the evolution of EM's philosophical understanding, and articulate a driving principle, consistent with its evolving interpretations.

4 Coherence

The principle we put forward as the drive underpinning spatial behavior is **coherence**: a "good" coupling between the agent and the environment. The drive towards coherence is about moving from confusion to clarity, resolving epistemic tension, "making sense of", becoming "ready-to-hand". A moment of coherence is the moment of the workspace being organized, the address being read from the notebook, the Tetris piece rotated, the *aha* moment of realization. It is constitutive – about state construction rather than instrumental goals, relational – involving both the agent and the environment, and subjective – specific to a particular agent at a particular time.

Before giving a candidate formalism in the next section, we provide further intuition by articulating relevant philosophy and ideas.

4.1 Philosophy and cognitive science

Our notion of coherence is related to "father of phenomenology" Merleau-Ponty's notion of *maximum grip* (Merleau-Ponty, 1962). According to Dreyfus (1996): *As an account of skillful action, maximum grip means that we always tend to reduce a sense of disequilibrium. What is experienced as disequilibrium and equilibrium is of course subjective. [...] is simply the body's ability to reduce tension or, to put it another way, to complete gestalts.* The term "completing gestalts" can refer to resolving psychological "unfinished business" or completing cognitive and visual patterns.

Highly relevant are *enactive* theories of cognition. Maturana & Varela (1980) define cognition not as passive information processing but a "structural coupling" of agent and environment. They argue that a living system and its medium undergo a history of recurrent mutual interactions through which the organism's internal structure and its environment co-vary, maintaining an operational congruence that is necessary for the organism's survival. Instead of relying on internal representations of an objective outside world, successful operation and adaptation within this coupled domain make up cognition.

There is an important connection with *meaning*. In *The Logic of Sense*, Deleuze (1969) argues that sense and meaning hover on the boundary, the "surface" between physical bodies (the outside) and mental propositions (the inside). Unlike representationalism which treats meaning like a static dictionary, enactive theories of cognition view meaning as continuously created through the agent-environment coupling (Varela et al., 1991).

Simondon (1958) in his theories of individuation, articulates "transduction", which relates with the dynamics of our notion of coherence. It is a *metastable* dynamic process where a disparity or energetic tension actualizes into a new form, temporarily stable until a subsequent tension arises. Distinct from *homeostatic* theories that Simondon critiqued, the resolution is not about eliminating tension, but developing structures that integrate new tensions well.

We would also like to nod to Schaul et al. (2011)'s work on *coherence progress*, which, though unrelated to the notions here, is a rare prior appearance of the term in RL-adjacent literature.

On constitution vs. instrumentalism. Finally, before moving on to the formalism we highlight a useful distinction and framing. A constitutive principle *makes up* a system, meaning it is inseparable from the thing itself, while an instrumental principle *is a means to achieve a separate goal*. Constitution *makes a thing possible*. We postulate that in this way coherence makes successful instrumental (reward maximizing) behaviour possible, and the two drives are in a complementary relationship.

5 Introducing the imaginary

How are we to operationalize these ideas without sacrificing their essence? If coherence is simply an auxiliary reward to be maximized, we have not respected the nature of the problem, and have simply introduces another tradeoff. Thus we take a fundamentally different route, and propose a shift not simply in terms of the existing algorithmic ontology, but in the ontology itself.

We propose (1) to **shift the domain of operation from real to complex numbers**, and (2) to **define coherence as phase alignment between complex features and weights.**

5.1 Features and weights

We first highlight the structural relationship at the heart of our proposal: the linear combination between features and weights. Whether the features are learned with a neural network, or are fixed and linear themselves, this combination is central to RL with function approximation. **We propose to ground the agent-environment distinction precisely into this formula: features are what is "outside", how the environment appears, and weights are on the "inside", what the agent values.** For some features ϕ and an action a whose weight vector is w_a, the value writes as a dot product:

$$Q(\phi, a) = \underbrace{\phi}_{\substack{\text{features} \\ \text{“outside"} \\ \text{environment}}} \cdot \underbrace{w_a}_{\substack{\text{weights} \\ \text{“inside"} \\ \text{agent}}}$$

Conventionally, we associate "state" with just features, and weights with parameters of action. Here, in line with the rest of our reinterpretations, we invite attention to their *relationship*. In order to be able to consider this relationship in a richer way, we propose to shift the domain **from real to complex numbers**. This adds dynamics to their interaction that amount precisely to the state-action entanglement we have referred to before.

5.2 From scalars to waves

A single complex number $z = |z|e^{i\theta}$ simultaneously encodes magnitude/amplitude ($|z|$ – intensity, salience, energy) and *phase/angle* (θ – timing, orientation).

When both the agent's weights and the environment's features are complex, their interaction shifts from a scalar dot product to an interaction of wave phenomena. They can interfere, constructively or destructively, via their phases. Writing a feature $\phi = |\phi|e^{i\theta_\phi}$ and a weight $w = |w|e^{i\theta_w}$, and with * denoting the complex conjugate, we have the complex product:

$$w^*\phi = (|w|e^{-i\theta_w})(|\phi|e^{i\theta_\phi}) = |w||\phi|e^{i(\theta_\phi - \theta_w)} = |w||\phi| \left[\cos(\theta_\phi - \theta_w) + i \sin(\theta_\phi - \theta_w)\right]. \quad (1)$$

This formula contains the entire mechanical intuition of coherence.

Phase alignment, $\theta_w = \theta_\phi$**.** The phase difference vanishes. $\cos(0) = 1$ and $\sin(0) = 0$. The imaginary part disappears, leaving a real-valued signal $|w||\phi|$.

Epistemic torque, $\theta_w \neq \theta_\phi$**.** When phases misalign, the $\sin(\theta_\phi - \theta_w)$ term activates. This leftover imaginary component represents epistemic torque – the rotational force or noise pushing the agent out of sync.

Conjugate symmetry, $w = \phi^*$**.** While phase alignment ensures a real-valued signal, full conjugate symmetry where $|w| = |\phi|$ creates a true *matched filter*. This structural mirror-symmetry fully absorbs environment fluctuations, neutralising the imaginary components that manifest as phase scatter (environment noise) and cognitive friction (variance of learning updates).

We thus propose to formalize *coherence as phase alignment*.[3] When phases align, the interaction transitions from destructive interference (noise/friction) to constructive interference (maximum energy transfer). This maximizes the real-valued signal, and produces a *resonant coupling*, two waves amplifying each other.

5.3 Complex RL

Let us now dive into the implications of this proposal for RL. We emphasize that our goal is not to propose precise algorithms, but to establish an ontological shift. By expanding the domain to complex numbers and defining coherence as phase alignment, we provide a conceptual and mathematical language that opens up new classes of algorithms and architectures.

5.3.1 Value

When the features and weights are complex, so is the value function. The relationship between the imaginary and the real from Eq. (1) reframes the nature of value in RL. Value no longer results only from maximizing reward, but from synchronising the agent-environment system.

The agent is still interested in maximizing real value. The imaginary part stores how much epistemic tension there is in the system, how "confused" the agent is. Actions can now result in increasing real value (temporal, pragmatic) or resolving the imaginary component (spatial, epistemic). Importantly, in this formulation, if the latter is successful, it results in a boost of the real value (which is maximized when w and ϕ are phase-aligned). Thus, the agent is *rewarded for the workspace being organized*, addressing the credit assignment problem from Sec. 2.1, but the increase in value comes from structural alignment rather than reward. This different source of value increase maps precisely onto the **distinction between instrumental and constitutive value**: instrumental values lead causally to outcomes, whereas constitutive values arise from intrinsic organization.[4]

5.3.2 The infinite game

There is a crucial difference between seeking coherence via phase alignment, which converts the imaginary into the real, and simply *minimizing* the imaginary. The latter is a form of avoiding surprise (e.g., Friston, 2010), which in its unconstrained form produces passive "dark room" agents. Because a coherence-oriented agent actively derives value from resolving incoherence into alignment, the imaginary is not treated as error, but as fuel to be realized.

Unlike standard RL, which is about convergence and "solving" the environment, this dynamic, like transduction, is about a *continuous integration of tension*. Provided that the environment continues to be a partner in it,[5] coherence thus characterizes the infinite game (Carse, 1986).

5.3.3 Generalized Markov

In physical systems, phase difference between interacting waves stores relative timing. The imaginary term of the value can thus be interpreted as *out-of-phase memory*. When it is zero, the coupled system does not require history of phase; the interaction becomes instantaneously localized in time.

[3] For the vector-valued case, which simply amounts to a per-component sum of Eq. (1): $\sum_j w_j^* \phi_j$, complete phase alignment may be too strong, and there are standard coefficients that measure the *degree of alignment*: from 0 (total incoherence) to 1 (perfect alignment).

[4] This naturally disturbs the usual Bellman mechanics, which anyway are not technically applicable to non-Markov non-stationary representations; integrating these dynamics is a key area for future development of principled algorithms.

[5] It is interesting to qualify when this is the case – clearly so for natural environments, that are not only vast but increasing in entropy, this is about *"big worlds"* that do not exhaust offering epistemic tension (Javed & Sutton, 2024).

This is precisely what qualifies a Markov interaction. So, instead of a Markov clock running uniformly at every step, the agent experiences *marked moments of relational coherence* whenever the coupled imaginary component vanishes.

Hypothesis: *Coherence serves as a relational generalization of the Markov assumption.*

Instead of "objective" Markov predictability of the future from the present state, the environment becomes predictable "subjectively", from the agent's perspective, because agent and environment are phase-aligned and "move together."

5.3.4 The boundary

There is freedom in choosing which agent-environment boundary to consider (Jiang, 2019; Harutyunyan, 2020). By ϕ we can consider anything from raw observation to intermediate environment features to the entire constructed state, and by w – anything from the entire representation to agent features to value, accordingly.

We notice that since observations remain real-valued, *no weights can achieve phase alignment with them in non-Markov spaces* – there will be irreducible epistemic torque. To be coherent, the agent thus *must* project observations into subjective complex-valued features, learn to understand the environment in a new way. *Pursuit of coherence thus necessitates development of representation,* and the agent has two paths: it can either learn to respond to the environment via its weights, or adjust the way it sees the environment, via its features.

5.3.5 Extended mind

The boundary seamlessly illustrates the EM thesis: *Whether it is an external artifact like a notebook in feature space ϕ, or biological memory in the agent state w, is irrelevant for the coupled coherence target.* Arguably, this addresses the coupling-constitution objection – whether the notebook constitutes or merely couples with cognition is irrelevant, because agent-side "cognition" w and notebook "reality" ϕ contribute to the coherence target interchangeably.

Now consider conjugate symmetry, the limit of resonance. It represents a complete mirror-like lock-and-key coupling between agent and environment. The environment becomes fully "transparent" to the agent, essentially part of the agent. Indeed, Eq. (1) reduces to $|w|^2$, thus literally erasing the boundary between the agent and the environment. We intuit that this is a form of the *stable complementary coupling* evoked by the extended mind, a strong integration of an external tool.

Finally, the overall ongoing interaction of the imaginaries at the boundary offers a computational metaphor for "sense-making" evoked by the third-wave view.

5.4 Bridges

The term "imaginary" was coined by Descartes in 1637 as a "derogatory label" for roots of equations that could not be represented on a 1D number line; he considered complex numbers useless, because they did not indicate physical, measurable quantities in the way real numbers did. Gottfried Leibniz on the other hand marvelled and referred to them as *a vital, "amphibian" bridge between being and non-being* (Gerhardt, 1850–1863) – noting that algebraists were frequently forced to pass through imaginary intermediates to arrive at real-valued solutions. A pivotal insight came from Gauss (1831) who introduced their *spatial* interpretation – conceiving of a complex number as a vector in the 2D plane, with i acting as a tangible $\pi/2$ rotation operator, and not an impossible quantity.

Moving from pure magnitudes to vectors with orientation adds a great deal of philosophical interpretive power. Orientation allows for intentionality – z is no longer just an "intensity", but is also pointing to something, is "about" something. The imaginary in general offers a natural store for the "ineffable" epistemic dimensions of experience (Lacan, 1977).

In AC physics, the imaginary represents *potential*: components that store energy to release later, while the real part is kinetic: components that immediately dissipate the energy exerted. Organizing a workspace or taking notes store "structural potential" for the future – whereas pragmatic instrumental actions immediately "realize" their work in the environment, advancing the agent towards the goal.

In control theory, the *Nyquist stability criterion* evaluates whether a closed-loop feedback system remains stable or degrades into wild oscillations. Rather than maximizing gain at all costs, stability requires managing phase lag across the feedback path to prevent error-canceling negative feedback from shifting into runaway, constructive positive feedback. This resembles the mathematics of conjugate symmetry, where phase alignment eliminates epistemic oscillations to enable reliable, resonant control.

Even in machine learning, phase alignment has recently emerged in some unexpected contexts. When analyzing the mechanics of *grokking* (Gromov, 2023) on modular arithmetic, the real-valued network was found to implicitly construct complex representations and *match their phase* to the periodicity of the task, which marked the shift from memorization to successful generalization – a coherence moment. *Rotary position embeddings (RoPE)* (Su et al., 2021) explicitly interpret 2D feature subspaces as complex planes, rotating query and key vectors by phase angles to encode relative information – offering promising empirical precedent for our ideas.

6 Closing

This work is entirely focused on the philosophico-technical argument, the proposal of coherence as a key principle of intelligent behaviour in non-Markov spaces, and the introduction of the imaginary as the vehicle of realizing it; leaving the full technical developments for future work. We believe this is an extremely promising path to the next generation of RL agents that go beyond gain-seeking logic, and are wired relationally. We hope we have provided some coherence to these views, and point to a few areas of future interest that have not been highlighted so far.

Convergence dynamics. RL with function approximation famously suffers from instability, in part due to a clunky projection step back into the feature space. Complex representations offer an entirely new dynamic to approach this, by offering an opportunity to reframe what it means to suitably project, taking the imaginary into account. In the control case, the reconsideration is in fact necessary, as complex numbers are not ordered – but it is also fortunate, since the $\max$ is a key source of instability.

Plasticity-empowerment "virtuous" cycle. We hypothesize that spatial action / extension create a "virtuous cycle" with the environment, where the agent is able to continuously increase its empowerment and plasticity (Abel et al., 2025) through the collaboration with the environment, and that this dynamic may be at the heart of a successful "infinite game."

Continual learning, exploration. Considering the classical RL open problems through the lens developed here, may provide novel insight. The metastable nature of coherence naturally lends itself well to the continual learning problem. There is a connection between epistemic torque and exploration.

Meaning beyond periodic phenomena. Our physics analogies are particularly well-suited to periodic phenomena (of which there are a relevant many). We hypothesize that the mechanisms of resonance and symmetry in the context of agent-environment relationship go beyond periodicity, and have the *potential* to represent enactive meaning.

Acknowledgments We thank David Abel, Mark Rowland, Tom Schaul, Hamza Merzić, Chen Yan, Hado van Hasselt, for thoughtful engagement with various stages of this work, the anonymous reviewers for open-minded feedback, and the Agency team at GDM for ever-insightful discussion, and an atmosphere of intellectual curiousity.

References

D. Abel, M. Bowling, A. Barreto, W. Dabney, S. Dong, S. Hansen, A. Harutyunyan, C. Khetarpal, C. Lyle, R. Pascanu, G. Piliouras, D. Precup, J. Richens, M. Rowland, T. Schaul, and S. Singh. Plasticity as the mirror of empowerment. *arXiv preprint arXiv:2505.10361*, 2025.

Fred Adams and Ken Aizawa. The bounds of cognition. *Philosophical Psychology*, 14(1):43–64, 2001.

Fred Adams and Ken Aizawa. Defending the bounds of cognition. In Richard Menary (ed.), *The Extended Mind*, pp. 67–80. The MIT Press, 2010.

Frederick Adams. Why we still need a mark of the cognitive. *Cognitive Systems Research*, 11(4): 324–331, 2010. Special Issue on Extended Mind.

Gloria Andrada. Epistemic complementarity: Steps to a second wave extended epistemology. In Inês Hipólito, Robert William Clowes, and Klaus Gärtner (eds.), *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artefacts*, pp. 253–274. Springer Verlag, 2021.

James P. Carse. *Finite and Infinite Games*. Simon & Schuster, 1986.

Andy Clark. Coupling, constitution and the cognitive kind: A reply to Adams and Aizawa. In Richard Menary (ed.), *The Extended Mind*, pp. 81–100. MIT Press, 2010.

Andy Clark and David Chalmers. The extended mind. *Analysis*, 58(1):7–19, 1998.

Gilles Deleuze. *Logic of Sense*. Columbia University Press, New York, 1969.

René Descartes. *La Géométrie*. De l'imprimerie de Ian Maire, 1637.

René Descartes. *Meditations on First Philosophy: In which the existence of God and the distinction between the human soul and body are demonstrated*. Paris, 1641.

Ben Deverett, Ryan Faulkner, Meire Fortunato, Greg Wayne, and Joel Z. Leibo. Interval timing in deep reinforcement learning agents. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Hubert L. Dreyfus. The current relevance of Merleau-Ponty's phenomenology of embodiment. *Electronic Journal of Analytic Philosophy*, 1996.

Chrisantha Fernando, Simon Osindero, and Dylan Banarse. The origin and function of external representations. *Adaptive Behavior*, 32(6):515–549, 2024.

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

Shaun Gallagher. The socially extended mind. *Cognitive Systems Research*, 2377, 2013.

Carl Friedrich Gauss. Theoria residuorum biquadraticorum, commentatio secunda. *Göttingische gelehrte Anzeigen*, 1831.

Karl Immanuel Gerhardt. *Leibnizens mathematische Schriften*. A. Asher & Comp. (Vol. 1-2) / H.W. Schmidt (Vol. 3-7), 1850–1863.

Sam J. Gilbert. Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory. *Cognition*, 247:105783, 2024.

Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7 (2):217–229, 2015.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

Anna Harutyunyan. What is an agent? *Unpublished*, 2020. URL http://anna.harutyunyan.net/wp-content/uploads/2020/09/What_is_an_agent.pdf.

Martin Heidegger. *Being and Time*. Harper and Row, 1927.

Khurram Javed and Richard S. Sutton. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: RLC 2024 Workshop*, 2024.

Nan Jiang. On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*, 2019.

Aaron Kagan and Charles Lassiter. The coupling-constitution fallacy: Much ado about nothing. *Pragmatics and Cognition*, 21, 2013.

Michael David Kirchhoff and Julian Kiverstein. *Extended Consciousness and Predictive Processing: A Third Wave View*. Routledge, 2018.

David Kirsh. The intelligent use of space. *Artificial Intelligence*, 73(1):31–68, 1995. Computational Research on Interaction and Agency, Part 2.

David Kirsh and Paul P. Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18:513–549, 1994.

Jacques Lacan. *Écrits: A Selection*. W. W. Norton & Co., 1977.

H. R. Maturana and F. J. Varela. *Autopoiesis and Cognition: The Realization of the Living*, volume 42 of *Boston Studies in the Philosophy of Science*. D. Reidel Publishing Company, 2nd edition, 1980.

Richard Menary (ed.). *The Extended Mind*. MIT Press, Cambridge, MA, USA, 2010.

Maurice Merleau-Ponty. *Phenomenology of Perception*. Routledge and Kegan Paul, 1962.

Rolf Pfeifer and Josh Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, Cambridge, MA, 2006.

Giulia Piredda. The mark of the cognitive and the coupling-constitution fallacy: A defense of the extended mind hypothesis. *Frontiers in Psychology*, 8:20–61, 2017.

Timotej Prosen. A moving boundary, a plastic core: A contribution to the third wave of extended-mind research. *Constructivist Foundations*, 17(3):220–230, 2022.

Evan F. Risko and Sam J. Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.

Tom Schaul, Leo Pape, Tobias Glasmachers, Vincent Graziano, and Jürgen Schmidhuber. Coherence progress: A measure of interestingness based on fixed compressors. In *Proceedings of the Fourth Conference on Artificial General Intelligence (AGI 2011)*, volume 6830 of *Lecture Notes in Computer Science*, pp. 21–30. Springer, 2011.

Gila Sher. *Epistemic Friction: An Essay on Knowledge, Truth, and Logic*. Oxford University Press, 2016.

David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Gilbert Simondon. *L'individuation à la lumière des notions de forme et information*. Presses Universitaires de France, 1958.

Jianlin Su, Yu Lu, Bo Pan, Ahmed Murtadha, Bo Wen, and Jun Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

John Sutton. Exograms and interdisciplinarity: history, the extended mind, and the civilizing process. In Richard Menary (ed.), *The Extended Mind*, pp. 189–225. The MIT Press, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Evan Thompson and Mog Stapleton. Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, 28:23–30, 2009.

F Varela, Evan Thompson, and E. Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.