Action control, forward models and expected rewards : representations in reinforcement learning

Rusanen, Anna-Mari

2021

Rusanen , A-M , Lappi , O , Kuokkanen , J & Pekkanen , J 2021 , ' Action control, forward models and expected rewards : representations in reinforcement learning ' , Synthese , no. þÿ 1 9 9 , p p . 1 4 0 1 7 1 4 0 3 3 . h t t p s : / / d o i . o r g / 1 0 . 1 0 0 7 / s 1 1 2 2 9 - 0 2 1 - 0 3 4 0 8 - w

http://hdl.handle.net/10138/352236
https://doi.org/10.1007/s11229-021-03408-w

Synthese (2021) 199:14017–14033
https://doi.org/10.1007/s11229-021-03408-w

ORIGINAL RESEARCH

Action control, forward models and expected rewards: representations in reinforcement learning

Anna-Mari Rusanen[1] · Otto Lappi[1] · Jesse Kuokkanen[1] · Jami Pekkanen[1]

Received: 1 June 2017 / Accepted: 7 September 2021 / Published online: 1 November 2021

Abstract

The fundamental cognitive problem for active organisms is to decide what to do next in a changing environment. In this article, we analyze motor and action control in computational models that utilize reinforcement learning (RL) algorithms. In reinforcement learning, action control is governed by an action selection policy that maximizes the expected future reward in light of a predictive world model. In this paper we argue that RL provides a way to explicate the so-called action-oriented views of cognitive systems in representational terms.

Keywords Representation · Reinforcement learning · Action control · Radical enactivism · Cognitive science

"…I see the new leaves on the tree
soon they are dancing with the whirlwind."
- Sonata Arctica, "Whirlwind"

This article belongs to the topical collection "Neuroscience and Its Philosophy", edited by Gualtiero Piccinini.

We thank all anonymous referees of this paper for their comments. Especially we thank the referee no. 3 for intensive, constructive and thoughtful feedback on several versions of the manuscript. Without this feedback this paper would have been far less than it turned out to be.

✉ Anna-Mari Rusanen
anna-mari.rusanen@helsinki.fi

Otto Lappi
otto.lappi@helsinki.fi

Jesse Kuokkanen
jesse.kuokkanen@helsinki.fi

Jami Pekkanen
jami.pekkanen@helsinki.fi

[1] Cognitive Science, Department of Digital Humanities, University of Helsinki, PO BOX 59, 00014 Helsinki, Finland

1 Introduction

The fundamental cognitive problem for active organisms is to decide what to do next in changing environments. According to representationalists, neurocognitive systems rely on representations when solving this problem. These internal model-like states allow organisms to plan and control sequences of behavior.[1] They enable adaptive, flexible and goal-directed action.

According to critics, however, the appeal to representations distracts rather than guides the research of cognitive phenomena (Chemero & Silberstein, 2008; Stepp et al., 2011; van Gelder & Port, 1995). For some, the representationalist framework is a result of a misinterpretation concerning what the experimental research actually entails. And some, such as recent radical enactivists, argue there is no satisfactory account of how contentful representational states drive action (Hutto & Myin, 2020). Instead, enactivists demand, we should explain action in terms of reactivations and re-enactments (Hutto & Myin, 2020).[2]

In their arguments, *sensorimotor action control* is typically taken as a paradigmatic example of a non-representational phenomenon. For example, Myin and Hutto (2015, p. 62) write "*… acts of perceptual, motor, or perceptuomotor cognition—chasing and grasping a swirling leaf—are directed towards worldly objects and states of affairs, or aspects thereof, yet without representing them*" (2015, p. 62). Cognitively, for a small and simple creature, such as a fly with little inertia, short temporal delays and few action choices, non-representational "reactivations" and "directedness" may suffice. For a human-sized complex organism with multiple possible actions, they may not.

When a human grasps a swirling object, a cognitive system must integrate information across multiple sensory sources (vision, touch, kinesthesia). It must also control multiple effectors (eyes, limbs, posture) in a goal-directed, temporally organised and purposeful way (Fiehler et al., 2019; Mischiati et al., 2015; Wolpert et al., 2011).[3] Furthermore, the environments where agents operate are constantly changing. To obtain success in such circumstances, agents can't rely only on actions that have been effective in the past. Instead, agents nust also explore options they have not tried before. Thus, they anticipate, prepare and plan. This may require surprisingly sophisticated cognitive synchronization, coordination and prediction abilities (Fiehler et al., 2019; Hayhoe, 2017).

In this paper, we analyse how reinforcement learning (RL) is used to study action control. RL is a computational framework in which the focus is on automating goal-directed learning and decision-making.[4] The heart of RL is an elegant and efficient trial and error algorithm. Its goal is to learn an optimal *action policy,* which maximizes the *expected rewards* in a given environment. The RL algorithm learns by exploring the possible actions and by observing the consequences of them.

RL-based action control models are more sophisticated than, say, simple proportional feedback models of the control theory from 1960s. Today, RL has an impressive

[1] Craik (1943): Ryder 2009, Grush (1997, 2004), Egan (2014, 2020).

[2] For discussion see Milkowski (2015), Zhao and Warren (2015).

[3] See also Wolpert and Kawato (1998), Wolpert and Ghahramani (2000), Desmurget and Grafton (2000).

[4] For a general account on RL-algorithm, see the work by Sutton and Barto (2018).

range of successful applications in AI, robotics and cognitive sciences. It is deployed, for example, in controlling the movements of robotic arms, navigating the routes in autonomous vehicles, predicting stock markets and playing various games. In computational neuro- and cognitive sciences, the framework is used to study various forms action and motor control, memory, decision-making and learning, and it is supported by a growing number of empirical and theoretical evidence.[5]

Crucially, RL-based action control is typically given a representational interpretation. It, however, differs from the portrait of representationalism drawn by recent enactivists. In enactivistic arguments, the paradigmatic case of a cognitive representation is a "percept", a sensory-like state that is used to "represent how things are with the world" (Hutto & Myin, 2020). The task of representations is to provide *information* about the environment, to which they are connected via sensory contact (Hutto, 2015).

In RL, the goal of the algorithm is to detect opportunities for successful action, not to construct descriptions of the world. The algorithm's computational objective is to learn the best possible action policy, in light of reward maximization. For doing so, it utilizes representational states. These states, however, are not "percepts".[6] Instead, they are goal-directed representations, which stand in for the estimated outcomes of action. Thus, this framework challenges the enactivist presuppositions on what action control representations are, and what they are used for.

2 Basics of reinforcement learning

Reinforcement learning (RL) is a computational framework, which focuses on automating goal-directed learning and decision-making.[7] Historically, RL is based on the work of early behaviorists. Thorndike's (1911) "Law of Effect" described how reinforcing events (i.e., reward and punishment) affect the tendency to select actions. For Pavlov (1927) and Skinner (1937), Thorndike's law provided a scientifically acceptable description for the mechanisms of conditioning. It offered a way to remove suspicious mentalistic concepts from psychology and neurosciences.

Later, Turing proposed an overall architecture for a "pleasure-pain system" in computers (Sutton & Barto, 2018). The actual formalisms of RL algorithm were formulated in the 1970s and 1980s, when computer scientists found a way to combine Optimal Control Theory, Temporal Difference Learning and Learning Automata.[8] Over the years, the approach has been developed further and complemented by a number of technical and conceptual additions.[9] As it stands, it is a family of efficient and sophisticated learning algorithms, which are widely used in computational cognitive and neurosciences, artificial intelligence and robotics.

[5] For overviews, see Eichenbaum and Cohen (2004), Niv (2009), Gershman and Daw (2017), Hayhoe (2017), Brea and Gerstner (2016).

[6] Rick Grush (1997) would call these percepts as "presentations".

[7] Sutton and Barto (2018).

[8] Sutton and Barto (2018).

[9] Just to mention some common supplements: The RL-algorithms utilize, for example, Markov decision processes (in the model-based RL), Q-learning (in the model-free RL), and Deep Neural Networks. For an overview of techniques, Sutton and Barto (2018).

In the 1990s, when cognitive neuroscientists began to utilize RL to model higher brain function, the initial "behavioristic" interpretation of reinforcement learning changed (Niv, 2009). Namely, one of their key theoretical insights was the idea that RL can be used to describe how neurocognitive systems learn by representing *value*.[10] Perhaps ironically, neuroscientists gave, thus, a thoroughly representational reinterpretation for a framework that was originally intended to eliminate such conceptualizations from the brain and behavioral sciences.

This reinterpretation of RL, however, does not fit well with recent enactivist views on representations. In enactivistic arguments, the task of representations is to provide information about the external environment (Hutto, 2015; Hutto & Myin, 2020). Representations are states, which (should) tell how the world is. In RL, however, the goal of the algorithm is not to detect the environment "in a correct way", but to learn the best possible action policy. That is, in RL, representations are goal-directed states.[11] Their success is not assessed by "veridicality" but by reward maximization.

To see this more clearly, we must unpack some core concepts of RL and take a look at a concrete example of a RL-based action planner.[12] Before that, some words of warning are in order. First, in this article we describe the algorithm in a simplified and vernacular way. Second, the concepts—such as *an agent*, *an environment, learning,* or *rewards*—are theoretical terms in computer sciences. They should not be confused with their counterparts in everyday language. For example, *an agent* does not refer to a whole organism, but a formal entity as specified in the formal description of the algorithm. Likewise, *an environment* is not a synonymous with the physical, concrete world around us. Instead, in RL "environment" denotes the so-called synthetic and technical environment, the *state space* of the RL model. It is the formally specified agent's world in which the algorithm interacts.

2.1 Core concepts of reinforcement learning

Generally, in RL, an *agent* learns[13] what to do by exploring the possible actions and by observing the consequences of its *actions*. The agent is not told which actions to

[10] One candidate for implementing this kind of predictive planning in the brain is the network linking the parietal cortex, frontal cortex, and the striatum. The parietal, premotor, and prefrontal cortices are most frequently reported as the areas activated in imagery of body movement as well as abstract cognitive operations (Deiber et al., 1998; Hanakawa et al., 2003; Sawamura et al., 2002). In combination with their interconnection with the cerebellum, those areas may store and update the predicted future states. The connections from those cortical areas to the striatum could be used for evaluation of the predicted states from hypothetical actions (Doya, 1999). A shortcut pathway from the cerebellum to the striatum through thalamus (Hoshi et al., 2005) may also be used for linking the internal models in the cerebellum and the value function in the striatum.

[11] See Anderson (2005) for the distinction between detection and goal-directed representations. See also Grush (1997) for a discussion on presentations and representations.

[12] For the purposes of this paper, we present a simplified account of RL where rewards are tied to states. Usually, the reward is associated with state-action pairs or state-action-next state tuples. However, this distinction is not relevant for the discussion at hand. For a more technical description, see the work by Sutton and Barto (1998, 2018).

[13] In this paper, we don't analyze learning itself but how a learned representation is used to control action.

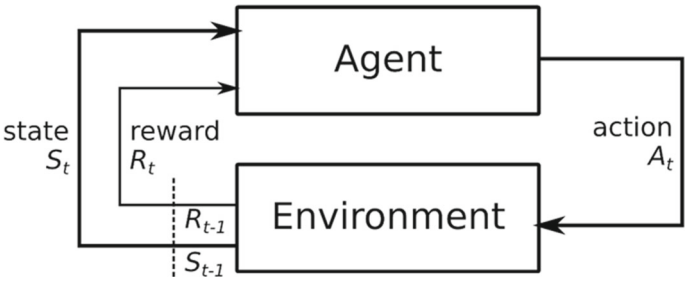

Fig. 1 The organization of a RL-algorithm

take. Instead, it must discover which actions yield the most *reward* in the long run by trying them out (Fig. 1).[14]

The goal of a RL algorithm is to learn an optimal *action policy* that maximizes the expected cumulative reward over time in a given environment. An action policy is, roughly, a strategy that an agent uses in pursuit of its goals. The policy dictates the actions that the agent takes as a function of (the agent's estimate of) the current *state*.

To learn the optimal action policy, RL algorithms must come to "know" the expected long-term cumulative reward, or *value,* of each state. Given estimates of each state's values and how actions affect state transition, the optimal action policy is to take the action that, on average, leads to the highest value in the next state: $a_t = \text{argmax}_a P(s_{t+1}|s_t,a,M')V(s_{t+1})$.[15]

A reward *r* is, roughly, the measure by which the immediate success or failure of an agent's actions can be estimated. A reward is a simple scalar, which can be negative (a punishment) or positive (a reward). Thus, a wide variety of entities can be described in terms of rewards. For instance, a reward signal can refer to different trade-offs, risk-seeking and risk aversing utilities, and many other combinations of objectives.

In RL, rewards are organism-dependent, and they are not properties of the physical world. Technically, a *reward function* R(s) is "a property of the organism", but the observed rewards are completely determined by the environment, and are not directly manipulable by the agent.[16] In other words, the reward function is, in a technical sense, "external" to the agent. The only way the agent can influence the rewards is through its actions.[17]

[14] There are different versions of RL algorithms. For overviews, see Kober et al. (2013), Sutton and Barto (2018).

[15] One-step prediction is sufficient for optimal behavior if the agent assumes that its value estimates perfectly reflect the environment. In practice and during learning, the value estimates are subject to uncertainty, and the prediction can be extended to multiple-step planning by recursively predicting further states.

[16] This prevents the agent from updating its reward signal—otherwise it could trivially maximize the reward by treating whatever happens as maximally rewarding. For adaptive behavior to emerge, the agent is not allowed to do this.

[17] Different agents may have different reward functions—one not more 'veridical' (or adaptive) than the other—even when the physical world they live in is the same (or even if the state representation for them is the same).

The concept of *value* refers to *the cumulative expected long-term reward*.[18] The amount of such a reward is specified by the value function: The value V(s) of a state is the expected future sum of a rewards (r_t) observed at time t, and future rewards that are typically *discounted* further in the future. The estimates of value take into account the following states, the reward accrued in those states, and their respective probabilities. For this reason, immediate high rewards do not always lead to maximal value. Or, a state might yield a low reward but still have a high value, if it is followed by other, high reward states.

2.2 Action planning in reinforcement learning

In the mainstream computational work on action, motor control is typically seen as mapping moment-by-moment control of movements in a simple motor task (e.g., hand movements, when grasping a leaf). The control is based on internal inverse and forward models (Miall & Wolpert, 1996; Weinsteiner & Botvinick, 2018; Wolpert et al., 1995). Inverse models allow to determine the motor commands necessary to achieve a desired state (for example, grasping a leaf), while forward models allow the system to predict the expected sensory feedback of a motor command.[19]

When the dynamics of action selection is approached in terms of RL,[20] the agent is thought to take an action—for example reaching a leaf—according to its action policy. The algorithm queries the model M′ for a state-action pair (s_t, a_t), and, in turn, receives the next state (s_{t+1}) and reward (r_{t+1}).[21] The algorithm receives a reward outcome in the form of a signal, and updates its action policy.[22] If the reward is positive, the algorithm strengthens its action policy (say, a sequence of motor movements to reach the leaf in a certain way). If the reward is negative, it weakens the policy (Fig. 2).

In contrast to many other computational approaches, in RL the algorithm does not only estimate the immediate next actions (as motor commands as sequences of physical movements), or focus only on what the expected sensory feedback (of a motor command) might be. Instead, it attempts to find an optimal policy through learning the values of actions at any state by estimating the expected future rewards. That is, it anticipates the success of actions *before* taking them. At the same time, it is able to estimate the success of the executed actions by receiving feedback in a form of rewards and punishments.

This combination of exploration and exploitation allows the systems to act in a more flexible, adaptive and proactive way. Moreover, these algorithms are able to respond to changes in the environment in a relatively efficient way. If the environment changes, the appropriate action becomes different. For instance, when the winter comes, instead

[18] For more technical details, see Sutton and Barto (2018).

[19] Typically the reaching dynamics is described as progressive adjustments of forward models to align the current observations on the feedback (Berniker & Kording, 2009; Haith & Krakauer, 2013).

[20] While in many other computational approach on motor processes the function of the forward model is to predict the sensory consequences of motor commands, in RL the function of the forward model is to *predict the amount of reward*.

[21] Thomas and Barto (2012) describes, how a goal-conditioned policy can be learned using multiple motor primitives.

[22] This is known as the basic form of RL. For more details, see Sutton and Barto (2018).

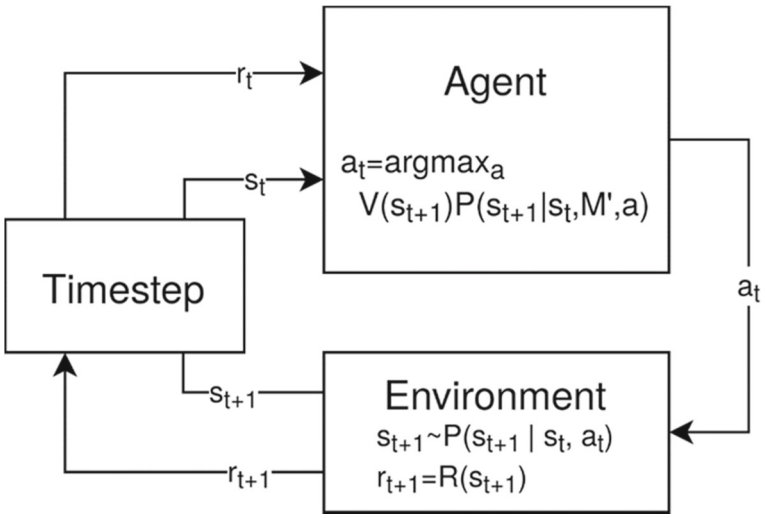

Fig. 2 Organization of value-based action selection in RL. At each timestep, the agent selects the action that produces maximal expected value in the next step

of leaves, snowflakes will swirl in the wind. Cognitively, this impacts on the goals of agent, too. The agent must be able to update its action policy, when necessary.

In RL, the algorithm can use *a forward model M′* for updating (Doya, 2008). In RL, forward models mimic the possible development of the environment, or more generally, they allow inferences to be made about how the environment will develop as a response to the agent's actions (in light of its action policy). Moreover, if such a forward model is available, with the so-called state transition rule P (new state|state, action), the agent can perform a following inference: If I take an action (a) in my current state (t), what is the next state (t + 1) I will end up in?

Technically, this planning and simulation procedure can be described as the maximization of the value up to a time T: $\sum$ T t = 0 rt + 1, where t indexes discrete time steps up to some maximum T, and r_t is the reward received at each step (Fig. 3).[23,24]

This allows that in RL, action planner systems are not limited to actions they've found effective in the past. They can also estimate the success of options that they have not tried before. Crucially, they are able evaluate which actions may yield the maximal results in the future. Thus, planner systems are able to *prepare* for the future, not only to learn from the past.

3 Action planner systems and representations

RL-based planner systems provide examples of how cognitive systems utilize *goal-directed* representations. First, the estimations of value can be taken to represent the

[23] Weinsteiner and Botvinick (2018).

[24] This approach differs to the so-called one-step model described in the Fig. 2.

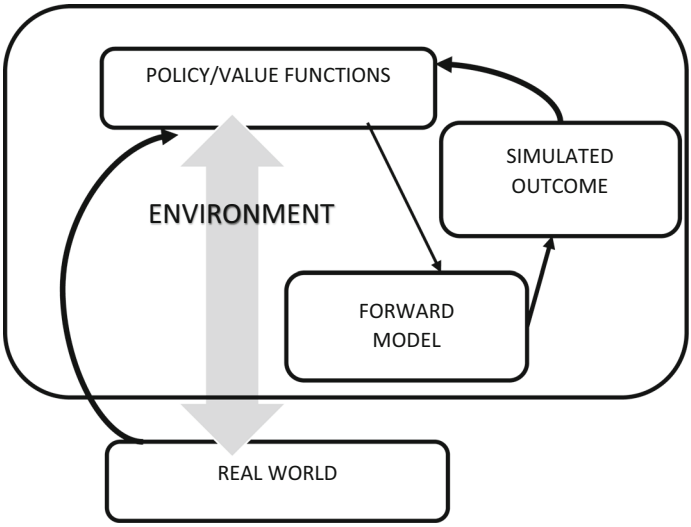

Fig. 3 Forward models and planning in action planners

goals in terms of expected outcomes (for example, the estimated amount of a long-term reward, if the agent grasps the leaf). Second, the action planners utilize forward models. They stand in for the future states of the algorithm's synthetic environment. That is, in RL the forward models do not represent the entities in the real world (say, hands grasping leaves), or the future trajectories of real world entities (say, the possible future trajectories of hands grasping new leaves). Instead, they represent the predicted states of the algorithm's environment in light of its action policy.

3.1 Forward models, synthetic environments and real world environments

In RL, the agent-environment construction is a part of the algorithm's specification, and the environment is literally a "synthetic" environment for the algorithm. Its content is specified in terms of the RL formalisms.[25] This synthetic environment can be seen as a surrogate, an abstract and idealized world model, which substitutes the real external, physical environment for the algorithm.[26]

In real-world tasks, however, the success of performance often requires a sufficient correspondence between the synthetic environment and the external physical environment. To systematically select appropriate and successful policies, the action planner system must take into account a significant number of external factors. For example, if the goal of a robot hand is to pick up a leaf in a real environment, relevant factors include the leaf's location, size, and distance to the hand, among other things.

[25] Grush (1997, 2004) provides a similar analysis in terms of emulators.

[26] One should not confuse this kind of representational surrogacy with ontological, causal or constitutive surrogacy. Something can ontologically or causally act as a surrogate for another thing without representing it, and an item can represent something as a surrogate without substituting ontologically (e.g. a hologram of Obi-Wan Kenobi as a surrogate of Obi-Wan in the Jedi high council).

The "sufficient correspondence", however, can be achieved in many ways.[27] Whether, and to what extent, the synthetic environment corresponds to the real-world environment depends, however, on the technical details of a particular application. Not all of them are representational. For example, in robotic reinforcement learning, the external environment can be designed to serve only as a source of feedback information. In this case, the control system receives information, uses it to update the parameters of the algorithms and to calibrate its actions. The feedback information, however, may play only a causal role. And, as Ramsey (2007) remarks, mere causal relations do not represent.

Or, applications can use "sensory-like" information as inputs. For example, in autonomous vehicles a variety of sensors (such as cameras, radars and lidars) can be utilized to scan the vehicle's environment.[28] Then, the sensor data can be combined, for instance, by using a computational technique called "sensor fusion".[29] The vehicle's control systems can take this fused data into account when they make the driving decisions. In this way, the sensor systems can "sense" or carry information about the relevant environmental factors, and the control systems can utilize this information, when necessary.[30]

The goal of a RL-based control unit is, however, to compute the driving decisions, not to track or represent the states of the environment in a correct or veridical way. The computational objective of RL algorithm is to maximize the reward, not to produce veridical representations of the environment.

Of course, one may program a RL application, such an intelligent camera, to mimic, say, perceptual processes and to learn world models in a more "representational" way (Silver et al., 2021).[31] For example, in a recent RL-based image recognition application the task is to learn a correct categorization for real-time video images (say, of leaves behaving in various ways)[32] During the learning process, the system receives *observations* (real-time videos, say, of swirling and non-swirling leaves) as input data.[33] Then, the system parametrizes the observations, and transforms them into hidden states.[34] At each step, the model predicts the policy, the value function, and the immediate reward. In RL, the hidden states are, thus, updated iteratively by a recurrent process that receives the previous hidden states and *hypothetical next actions* in the systems.

[27] Sutton and Barto (2018).

[28] For an overview of various sensor technologies, see Campbell et al. (2018).

[29] For technical details, see Yeong et al. (2021) for an overview.

[30] This is a rough, vernacular and superficial description. For example, see Yeong et al. (2021) for an overview.

[31] According to some hypothesis (Roefsama et al., 2010), in mammalian brains certain sensory systems might utilize RL-algorithms.

[32] For the original example, see Schrittwieser et al. (2019).

[33] The specialty of RL is that computationally they are not designed to produce the predictions about the statistical patterns of inputs (or similarities in data points) as their outcomes. *Instead, the algorithm´s computational goal is to maximize amount of long term reward in a light of an action policy.*

[34] In some current RL applications, the system builds a model of the world by predicting future observations given the past (Hafner et al., 2019).

As a result, as Schrittwieser et al. (2019) put it, "*there is no direct … requirement for the hidden state to capture all information necessary to reconstruct the original observation, nor is there any requirement for the hidden state to match the unknown, true state of the environment; nor any other constraints on the semantics of state… they are free to represent the environment in whatever way is relevant to maximize current and future rewards."* That is, the algorithm's goal is not to match the representational states of the system with the "true states" of the external environment. Or, its task is not to use (current or past) "observations" (about the external environment) to construct veridical descriptions of the objective world as such. Instead, the algorithm's job is to decide which actions it should take to maximize the reward. To do so, it is free to represent the states of physical environment in whatever way which is relevant for maximizing the rewards.

Furthermore, the specialty of action planner systems in RL is that they do not only use (current or past) "observations" (about the external environment) to predict future rewards. Instead, they *also* estimate the impact of future actions to predict future rewards. In RL, the *forward* models tell what the possible development of algorithm's synthetic environment may be. These future-oriented states allow the algorithm to estimate how its environment will develop as a response to agent's actions (in light of its action policy). Thus, these estimates have no existing target systems in the real world. Instead, they stand in for anticipated possible future states.[35] Hence, in RL, forward models represent "possible worlds", not real environments.

3.2 Value representations

Value estimations stand in for the total amount of reward an agent can expect to accumulate over the future. Their content is specified in terms of the value function and other formalisms of the algorithm. While a value is the farsighted judgement of what is good in the long run, a reward refers to the measure by which the immediate success of an action can be estimated. It indicates the intrinsic short-term desirability of environmental states for the agent.[36]

Rewards, thus, are not signals that indicate the presence of states or entities per se. A reinforcing stimulus (physical feature) only presents an organism with reward or value depending on its reward function. The aim of the reward function is not to represent whether the stimulus is "really" valuable or rewarding.

In animal experiments, of course, rewards are often operationalized as food or drink. For example, a capuchin monkey can be taught to do various tasks by rewarding it with a treat, such as a grapefruit (Brosnan & de Waal, 2003). Still, despite this operationalization, the reward itself is not defined as the grapefruit (or its glucose).

[35] One might be tempted to think that if M stands in for E, and if E stands in for W, then M stands in for W. However, this would violate the so-called intransitivity condition of representational relations (Goodman, 1976). Thus, even in the case, in which the synthetic environment represents the real world environment, the forward model represents the synthetic environment, and the synthetic environment represents the real world environment.

[36] In RL framework, the reward cannot be normatively evaluated in terms of satisfaction conditions that the reward signal would be 'about'.

Instead, the reward is like the "pleasure" related for getting such a thing (Sutton & Barto, 2018).

Moreover, *rewards are not reducible to an agent-independent physical property or object*. They are organism-dependent features, not organism-independent causes of neural responses. Factors, such as the hunger-state of the organism, impact the quantity and the quality of rewards. Sugar, for example, is not equally rewarding for every organism (with different reward functions). For a capuchin monkey, it usually is. For a cat, it isn't. How rewarding sugar is depends on the current glucose metabolism, or generally the physiological state of the organism. Sugar is more rewarding when one is hungry, and less rewarding when one is not.

4 Fly detectors and goal-directed representations

The value estimations and forward models are examples of goal-directed representations. Their role is to detect opportunities for successful action, not to construct veridical models of the objective world as such. Thus, they do not fit well with the recent enactivistic arguments, where the paradigmatic case of a cognitive representation is a "percept", a mental state used to "represent how things are with the world" (Hutto & Myin, 2020). These representations, as Hutto (2015) puts it, are connected with the world via "sensory contact".

From a cognitive science point of view, Hutto and Myin's arguments continue the legacy of "fly detectors".[37] This legacy dates back to receptive field studies from the late 1950s (Hubel & Wiesel, 1959; Lettvin et al., 1959). In these early studies on sensory mechanisms, the focus was on signal transformation properties of frog ganglion cells, later known as "fly detectors" (Lettvin et al., 1959). The ganglion cells were found to respond to small, black, fly-like dots moving against a stationary background in the frog's visual field. A few years later Hubel and Wiesel (1962) proposed a way in which "pooling mechanisms" might explain how the cells of mammalian visual cortex are able to detect more complex features by combining simple responses.

Under the influence of fly detectors, the experimental research of sensory processes focused on a bottom-up feature detection for decades. This framework impacted also on philosophers, and fly detectors dominated the discussion on representations for a long time. For example, in the 1980s, the analysis of representations focused almost completely on the questions of (1) whether a representation of a fly is really *about* flies, (2) how to make the leap from the signal transformation properties of, say, ganglion cells into semantic properties of receptors, or (3) of how to specify the *content* determination of these representations in a satisfactory naturalistic way (Dretske, 1981; Fodor, 1992; Millikan, 1989).

From a cognitive point of view, fly detectors are stimulus-based representations. That is, the activation of a detector representation is taken to require a causal association with preceding stimuli, a "neural" signal that triggers the representation (or causes an indicator to fire). For example, in the crude causal theory a stimulus—say, a black dot in the visual field of a frog—is a proximal cause for the activation or the triggering

[37] See also Grush (1997), Anderson (2005).

of the tokening of a representation "fly". The stimulus is caused by a signal, and the external source of a signal—an entity, say, a fly or a mosquito—is assumed to exist in the physical environment. In teleosemantics or indicator semantics, the stimulus is typically taken to be responsible, for example, for the firing of indicator mechanisms, or to play a part in the causal specification of an indicator mechanism.[38]

In stimulus-based representations, the source of a signal (causing the stimulus or the source for a cause that is responsible the firing of an indicator) is taken to exist, somehow, as a distal object in the physical environment (including, possibly, parts of the organism's body). Obviously, RL-based action planning representations cannot be specified in such terms. They are not caused or triggered by stimuli. For example, in value representations the content of a reward is specified by the reward function. Stimulus is, however, not a part of the reward function. Rewards, simply, aren't rewarding because they are triggered by entities that happen to function as reinforcers for the agent. Instead, in RL, reinforcers reinforce *because* they elicit reward.

In value or reward representations, thus, the fly-detector based formulation on signals and stimuli is turned on its head: in RL, the signal is not "a reward signal", because it triggers or causes a reward stimulus. Instead, a stimulus is "rewarding" because *it causes* a reward signal (i.e., signal that acts the way reward acts in RL). Furthermore, the reward signal itself does not indicate the presence of some independent feature of the environment (such as the presence of, say, glucose in a grapefruit as a reinforcing stimulus). Although the reward can be operationalized as a concrete item (such as treats, say grapefruits, in animal experiments), the reward itself is not the grapefruit, but the feeling of "pleasure" related to it. Thus, the question of whether the semantic content of a reward signal is *really* glucose, grapefruit, fly or whatever, simply does not arise.

To sum up, neither rewards nor value representations (as estimated long-term rewards) can be specified in terms of "fly detectors", or any other framework, in reference to environmental stimuli.[39] Obviously, these organism-dependent and goal-oriented representations raise very difficult problems of neural encodings of their

[38] In philosophical theories, the role of stimulus is specified in multiple ways. In *causal information semantics*, the proximal stimulus (in the visual field) carry information about the distal entity. Since the stimulus is also caused by the distal entity, in information semantics the stimulus it also has also a causal role in the tokening of a representation in information semantics. In *teleosemantics*, stimuli play a role, for example, in the causal specification for of the mechanisms that implement "teleological functions" in perceptual detectors or indicators (Dretske, 1995; Millikan, 1989). In some variants of teleosemantics, stimuli are causal components of the mechanism, which activates the detectors to fire (Neander, 2018). For example, if a certain perceptual state P of a frog is caused to token whenever a stimulus—small dark thing—moves in the visual field of the frog. For Millikan´s consumer teleosemantics (1989), the role of stimulus is, roughly, to activate perceptual state P. Then, this state causes its 'consumer' in the system to initiate a sequence of operations that result in the frog lashing out its tongue. For the producer-teleosemantics, the function of producers is to respond to visual features of bugs by "producing" representational states. This makes P carry information about the stimulus i.e. the visual features of bugs. For instance, Neander (2018, p. 146) writes, "sensory-perceptual representation of type R has the function to carry information about stimuli of type C is to say that the system that produces Rs has the function to produce them (i.e., change into R-states) in response to C-type stimuli.".

[39] The "content" of reward is specified in terms of reward functions, and it is dependent on the description of a particular organism *as* a RL agent. That is, the physical properties in the environment that constitute the reward for a specific organism are not given by any organism-independent relevance, veridicality or

formal and abstract properties. As things stand, there is no appropriate philosophical account for them.

As we see it, the algorithmic methods of cognitive and computer sciences may provide some resources for analyzing them. For example, the formal descriptions of "synthetic environment" and "expected rewards" may provide a more fertile way to make an autopsy for these organism-derived, abstract representations than the traditional philosophical analysis in terms of "secondary qualities", "observer-dependent representations" or "organism-produced representations in absence of environmental targets". Computational methods provide not only tools for simulating these phenomena in a controlled way but, more importantly, they may also offer concepts for analyzing them in an exact and mathematically tractable way.

5 Diversity of representations

When assessing what is the most plausible empirical story for behaviors like "chasing and grasping a swirling leaf", one should remember that we live in a complex and changing world. In dynamic environments, even simple action—such as grasping a new swirling leaf—is a complicated challenge. To solve this challenge, cognitive systems do not only react, but also explore new strategies and learn from observing the consequences of their actions. Moreover, to select successful policies, they must take into account a significant number of external factors, and integrate information from several sources in a purposeful way. Thus, successful action requires information from multiple perspectives.

From a neurocognitive point of view, cognitive systems use different representations for different purposes.[40] Perceptual and sensory systems allow systems to detect the environment and to get feedback from the states of the body. Motor representations as motor commands help the agent to coordinate the behavior of body. They encode explicit instructions for sequences of physical movements (Miall & Wolpert, 1996; Mylopoulos & Pacherie, 2017; Wolpert et al., 1995). And, goal-directed representations are used to detect and estimate opportunities for successful action.

Representational states vary from neural coding in sensory cells to higher-order perceptual processing, and from on-line representations to memory-based representations of previous experiences. And, they vary from simple motor commands to complex metacognitive simulations of possible actions. While some of them, such as fly-detectors or simple encodings of visual or acoustic signals, are stimulus-based and track the external targets via sensory apparatus, other representations such as motor representations or goal-directed representations,[41] do not. And, while some

Footnote 39 continued
other criteria. Instead, the reward function is a description of aspects of the physiology and the cognition of a particular organism as such an agent.

[40] For similar remarks, see Grush (1997, 2004), Anderson (2005).

[41] According to Millikan (1984), *indicative representations* represent how the world is. They track the actual states of the world. *Imperative representations*, such as motor commands, represent how the world will, or should, be. For instance, motor representations stand in for the desired outcome of the sequence of physical movements. For contemporary discussion, see Thomson and Piccinini (2018), Pavese (2020).

representations are organism-independent, such as sensory encoding of simple physical quantities like the speed and direction of motion, others, such as estimates of long-term rewards, are organism-dependent.

Clearly, this diversity challenges the enactivist presupposition that to represent is to represent only as fly detectors, or generally, as sensory states do. Furthermore, it pushes us to ask: to what extent can these various representational states be analyzed *only* in terms of a framework which was originally developed to explore the signal transformation properties of a frog's ganglion cells. Perhaps we should let go of the assumption that only states that track external environment count as representational, and analyze the representational diversity as it is. Eventually, this should not be taken as a matter of philosophical convictions or presuppositions. Instead, it should be taken as a question of roles that representations play in explaining action control scientifically.

6 Conclusions

In this article, we have illustrated how reinforcement learning (RL) algorithms are used to study the cognitive dynamics of action control. When action control is characterized in this way, it is understood as the control system of a cognitive agent who can learn, anticipate and adapt.

In RL, the cognitive dynamics of action control is seen as forms of computationally specified learning-as-decision making processes (Sutton & Barto, 2018). RL provides an exact and algorithmic way to describe various aspects of these processes. Furthermore, this approach is widely and successfully used in simulating and analyzing them in computational cognitive sciences, artificial intelligence and robotics.

In RL, the goal of an algorithm is to maximize the long-term reward. For doing so, the algorithm utilizes internal representational states. These representational states, however, are not sensory-like "percepts". Instead, they are goal-directed. Thus, this algorithmic framework provides examples of representations which challenge the enactivist conviction that to represent is to represent as fly detectors do. Moreover, it impugns the enactivist intuition that action-oriented perspective implies the need for a non-representational theory of cognitive systems. Instead, in the light of RL, it implies a need for updating intuitions on what action control representations are, and what they are used for.

Funding Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

Anderson, M. L. (2005). Representation, evolution and embodiment. In D. Smith (Ed.), *Evolutionary biology and the central problems of cognitive science, a special issue of Theoria et Historia Scientiarum* (Vol. 9, No.1, pp. 1–18).

Berniker, M., & Kording, K. (2009). Estimating the sources of motor errors for adaptation and generalization. *Nature Neuroscience*, *11*, 1454–1461. https://doi.org/10.1038/nn.2229.

Brea, J., & Gerstner, W. (2016). Does computational neuroscience need new synaptic learning paradigms? *Current Opinion in Behavioral Sciences*, *11*, 61–66. https://doi.org/10.1016/j.cobeha.2016.05.012.

Brosnan, S., & de Waal, F. (2003). Monkeys reject unequal pay. *Nature, 425*, 297–299. https://doi.org/10.1038/nature01963

Campbell, S., O' Mahony, N., Krpalkova, L., Riordan, D., Walsh, J., Murphy, A., & Ryan, C. (2018). Sensor technology in autonomous vehicles: A review (pp. 1–4). https://doi.org/10.1109/ISSC.2018.8585340

Chemero, A., & Silberstein, M. (2008). After the philosophy of mind. *Philosophy of Science, 75*, 1–27.

Craik, K. (1943). *The nature of explanation*. Cambridge University Press.

Deiber, M. P., Ibañez, V., Honda, M., Sadato, N., Raman, R., & Hallett, M. (1998). Cerebral processes related to visuomotor imagery and generation of simple finger movements studied with positron emission tomography. *NeuroImage*, *7*(2), 73–85. https://doi.org/10.1006/nimg.1997.0314.

Desmurget, M., & Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. *Trends in Cognitive Sciences, 4*(11), 423–431.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks : The Official Journal of the International Neural Network Society*, *12*(7–8), 961–974. https://doi.org/10.1016/s0893-6080(99)00046-5.

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience, 11*(4), 410–416.

Dretske, F. (1981). *Knowledge and the flow of information*. MIT Press.

Dretske, F. (1995). *Naturalizing the mind*. MIT Press.

Eichenbaum, H., & Cohen, N. (2004). *From conditioning to conscious recollection: Memory systems of the brain*. Oxford University Press.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*, 115–135. https://doi.org/10.1007/s11098-013-0172-0.

Egan, F. (2020). A deflationary account of mental representation. What are mental representations? In J. Smortchkova, (Ed.), ISBN 978-0-19-068667-3 (pp. 79–100). Oxford University Press.

Fiehler, K., Brenner, E., & Spering, M. (2019). Prediction in goal-directed action. *Journal of Vision, 19*(9), 10.

Fodor, J. (1992). *A theory of content and other essays*. MIT Press.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*(1), 101–128.

Goodman, N. (1976). *Languages of art* (2nd ed.). Hackett.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology, 10*(1), 5–23. https://doi.org/10.1080/09515089708573201

Grush, R. (2004). The emulation theory of representation: motor control, imagery and perception. *Behavioral and Brain Sciences, 27*, 377442.

Hafner, D., Lillicrap, T., Ba, J., & Norouzi, M. (2019). Dream to control: Learning behaviors by latent imagination. arXiv:1912.01603 [cs.LG]

Haith, A. M., & Krakauer, J. W. (2013). Model-based and model-free mechanisms of human motor learning. *Advances in Experimental Medicine and Biology*, *782*, 1–21. https://doi.org/10.1007/978-1-4614-5465-6_1.

Hanakawa, T., Immisch, I., Toma, K., Dimyan, M. A., Van Gelderen, P., & Hallett, M. (2003). Functional properties of brain areas associated with motor execution and imagery. *Journal of Neurophysiology*, *89*(2), 989–1002. https://doi.org/10.1152/jn.00132.2002.

Hayhoe, M. M. (2017). Vision and action. *Annual Review of Vision Science, 3*, 389–413.

Hoshi, E., Tremblay, L., Féger, J., Carras, P. L., & Strick, P. L. (2005). The cerebellum communicates with the basal ganglia. *Nature Neuroscience*, *8*(11), 1491–1493. https://doi.org/10.1038/nn1544.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 124*(3), 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(45), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

Hutto, D. (2015). Overly enactive imagination? Radically re-imagining imagining. *The Southern Journal of Philosophy, 53*, 68–89. https://doi.org/10.1111/sjp.12122

Hutto, D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content*. MIT Press.

Hutto, D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT Press.

Hutto, D., & Myin, E. (2020). Deflating deflationism about mental representation. What are mental representations? In J. Smortchkova (Ed.), ISBN 978-0-19-068667-3 (pp. 79–100). Oxford University Press.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *International Journal of Robotics Research, 32*, 1238–1274.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE, 47*, 1940–1951.

Miall, R., & Wolpert, D. (1996). Forward models for physiological motor control. *Neural Networks, 9*, 1265–1279.

Miłkowski, M. (2015). The hard problem of content: Solved (long ago). *Studies in Logic, Grammar and Rhetoric, 41*(1), 73–88.

Millikan, R. (1984). *Language, thought, and other biological categories : New foundations for realism*. MIT Press.

Millikan, R. (1989). Biosemantics. *The Journal of Philosophy, 86*, 281–297.

Mischiati, M., Lin, H.-T., Herold, P., Imler, E., Olberg, R., & Leonardo, A. (2015). Internal models direct dragonfly interception steering. *Nature, 517*, 333–338. https://doi.org/10.1038/nature14045

Myin, E., & Hutto, D. (2015). REC: Just radical enough. *Studies in Logic, Grammar and Rhetori, 41*(1), 61–71.

Mylopoulos, M., & Pacherie, E. (2017). Intentions and motor representations: The interface challenge. *Review of Philosophy and Psychology, 8*(2), 317–336.

Neander, K. (2018). *A mark of the mental*. The MIT Press.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139–154.

Pavese, C. (2020). Practical representation. In E. Fridland, & C. Pavese (Eds.), *The Routledge handbook of philosophy of skill and expertise* (pp. 226–244). Routledge.

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford Univ. Press.

Ramsey, W. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511597954.

Roelfsema, P., Ooyen, A., & Watanabe, T. (2010). Perceptual learning rules based on reinforcers and attention. *Trends in Cognitive Sciences, 14*(2), 64–71.

Sawamura, H., Shima, K., & Tanji, J. (2002). Numerical representation for action in the parietal cortex of the monkey. *Nature*, *415*(6874), 918–922. https://doi.org/10.1038/415918a.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T. P., & Silver, D. (2019). Mastering Atari, Go, Chess and Shogi by planning with a learned model. arXiv:1911.08265

Silver, D., Singh, S., Precup, D., & Sutton, R. (2021). Reward is enough. *Artificial Intelligence, 299*, 2021.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Miller and Konorski. *Journal of General Psychology*, *16*, 272–279. https://doi.org/10.1080/00221309.1937.9917951.

Stepp, N., Chemero, A., & Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science*, *3*, 425–437. https://doi.org/10.1111/j.1756-8765.2011.01143.x.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT press.

Thomas, P., & Barto, A. (2012). Motor primitive discovery. *International Conference on Development and Learning - EpiRob (ICDL)*, 1–8. https://doi.org/10.1109/DevLrn.2012.6400845.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds & Machines, 28*, 191. https://doi.org/10.1007/s11023-018-9459-4

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Macmillan Press. https://doi.org/10.5962/bhl.title.55072.

van Gelder, T., & Port, R. (Eds.). (1995). *Mind as motion*. MIT Press.

Weinsteiner, A., & Botvinick, M. (2018). Structure learning in motor control: A deep reinforcement learning model. CoRR arXiv:1706.06827.

Wolpert, D., Ghahramani, Z., & Jordan, M. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880–1882.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Network*, *11*(7–8), 1317–1329. https://doi.org/10.1016/s0893-6080(98)00066-5.

Wolpert, D. M., Diedrichsen, J., & Randall Flanagan, J. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience, 12*(12), 739–751.

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience, 3*, 1212–1217.

Yeong, D. J., Velasco-Hernandez, G., Barry, J., & Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, *21*(6), 2140. https://doi.org/10.3390/s21062140.

Zhao, H., & Warren, W. H. (2015). On-line and model-based approaches to the visual control of action. *Vision Research, 110*(Part B), 190–202.