Reinforcement learning is a set of mathematical and computational tools for handling sequential decision problems. As a branch of economics and operations research, it is closely related to the study of choice behavior and theories of rational choice in particular. As a branch of computer science, it sits alongside supervised and unsupervised learning as one of the methodological pillars of machine machine learning. And as a branch of cognitive science, it stands as one of the best-confirmed computational analyses of animal behavior, unifying a wide range of observations and enjoying precisely identified neural correlates. Despite this impressive pedigree, reinforcement learning has garnered little attention from the philosophical community. This is a shame. Reinforcement learning can shed light on many questions of philosophical interest, and raises interesting questions of its own. I will indicate some of these questions below. I claim that reinforcement learning provides an account of the structure and content of certain mental representations, and that this account reveals surprising features of these representations. Before explaining this account in any greater detail, I will provide an overview of reinforcement learning and summarize some of the field's key findings.

1 RL and know-how

The nature of know-how has occupied a prominent place in contemporary philosophy of mind and action. Much of the literature focuses on a question raised by [CITE: Ryle]'s criticisms of *intellectualism*, the view that know-how centrally involves a kind of propositional knowledge. *Anti-intellectualists* (such as Ryle) deny this. My goal in this section [paper?] is to demonstrate that computational reinforcement learning provides a range of tools for probing the nature of knowledge-how. I will also argue that it provides compelling counterexamples to the intellectualist thesis. But I think that the former contribution is more important: our practical capacities exhibit rich and variegated structures; it matters more to understand these structures on their own terms than to decide whether know-how is, always and everywhere, a form of know-that. So, at any rate, I will argue. Still, the debate between intellectualists and anti-intellectualists provides a nice place to start, so I will begin by laying out the two positions. Then, I will introduce the basics of reinforcement learning, sharing technical details on a need-to-know basis. With those basics on the table, I will illustrate some central uses of reinforcement learning in the cognitive and neural sciences, with an eye to their bearing on the intellectualist position. After explaining why I take these cases to refute intellectualism and considering rejoinders, I go on to illustrate how computational concepts from reinforcement learning can shed light on various aspects of know-how, abilities, and skills, such as their relation to chunking and motor schemata, the distinction between habitual and goal-directed behavior, the question of automatization, and ...

2 Intellectualism and Anti-Intellectualism

In this section, I will endeavor to clarify the two central positions. This is not a trivial task. The literature has revolved around three concepts: *skill*, *ability*, and *know-how*, and their relation to propositional knowledge. Various authors seek to identify one or more of these phenomena. For example, [CITE: Noe], following [CITE: Ryle], maintains that know-how and ability are one. [CITE: Stanley and Williamson] deny this, and maintain instead that know-how is a species of propositional knowledge (which anti-intellectualists deny). Some, like [CITE: Stanley and Krakauer], identify skill and know-how but divorce them from abilities, holding that one can know how to do something without being able to do it. And so on. It is therefore difficult to characterize the subject matter in a theory-neutral way (that is, in a way which wouldn't draw objections from at least one party). Better to proceed by way of examples.

Examples of the phenomena in question include my knowing how to swim, how to write a philosophy paper, how to get to the grocery store, how to play a video game, how to count to 10, and how to play the bass. To these we can add that Alva Noë's dog knows how to catch a Frisbee [CITE: Noe, 289] and that trained rats know how to navigate out of a maze. (Note that even [CITE: Stanley and Williamson], the most promiment contemporary proponents of intellectualism, are happy to ascribe know-how to dogs.) The first thing that should strike us about this collection of examples is its diversity: it is *ex ante* implausible, I think, that all of these examples fall under the same mental kind. Indeed, as I shall argue below, they do not: although there are good reasons to lump them together in ordinary talk, they are underwritten by importantly different cognitive mechanisms. Ordinary talk is (here as elsewhere) no guide to mental organization. As such, I will talk of know-how, skills, and abilities more or less interchangeably, but without wishing to imply anythign about their identity or distinctness.

Be that as it may, here is the intellectualist thesis:

> Intellectualism: To know how to φ is to know, for some way w of doing φ, that w is a way of doing φ.

(φ ranges over actions.) Note that knowing that w is a way of doing φ is knowledge of a proposition: it entails holding a propositional attitude (belief) toward the proposition that w is a way to do φ. Many intellectualists add that the proposition in question must feature the way w under a *practical mode of presentation* [CITE: Stanley and Williamson, Pavese]. Practical modes of presentation are presented as a species of Fregean modes of presentation. Modes of presentation support a fine-grained notion of mental content of the kind suitable for psychological and rational explanation [CITE: Fodor, Burge, Rescorla]. Practical modes of presentation index representational content to the exercise of practical capacities. Although we do not consider this notion irredeemably obscure, our argument will not turn on it, and so we omit it from our discussion.

Intellectualism is, as stated, a very strong thesis. It implies not only that knowing how to do something requires a capacity for propositional thought, but

also the possession of propositional *knowledge*. Moreover, having such knowledge is not merely necessary for possessing the relevant know how. It is constitutive of (and hence sufficient for) know-how. Not all who call themselves "intellectualists" will sign on to the thesis in its full strength ([CITE: XXXX], for example, requires mere belief rather than knowledge, and at times [CITE: Stanley and Krakauer] seem to think that propositional knowledge is merely necessary for, but perhaps not constitutive of, know-how). But, again, our arguments below will not turn on these features of the view, so we can let it stand in its purest form.

It *is* important, however, to distinguish Intellectualism from a weaker thesis:

> Way-Representationalism: knowing how to φ requires (or consists in) having some representation, whether propositionally structured or not, of a way to do φ.

This weaker thesis does not require knowers-how to be capable of propositional thought, though we may add that the representation of the way must be through a practical mode of presentation. Although I will argue this view should also be rejected, not all considerations against Intellectualism will apply to Way-Representationalism.

Finally, we must also distinguish the two foregoing views from

> Representationalism: knowledge-how requires some representations or other.

This view is much weaker. I do not know whether it is true, but I will suggest some tools for assessing it below.

3 Reinforcement Learning

Broadly speaking, reinforcement learning models agents learning to do the best for themselves as they interact with their environment. I begin to sharpen this characterization by explaining reinforcement learning from a computational perspective, before discussing its use in cognitive science.

3.1 Reinforcement Learning in Computer Science

Reinforcement learning models decision-making problems that have a certain mathematical structure: Markov Decision Processes (MDPs). An MDP consists of an agent interacting with an *environment.* At any point, the environment is in one of a range of possible *states* and the agent can perform an *action* from a set of alternatives. The agent's action may affect the environment, resulting—perhaps probabilistically—in a *next state* and occasioning—perhaps probabilistically—a *reward*. The reward is modeled as a simple numerical signal. The agent's goal is to accummulate as much reward as possible over the course of its interaction

with its environment. More precisely, its goal is to maximize its *expected return*, defined as the expected value of its cumulative, discounted rewards. This expected return is a function of the environmental dynamics and of the agent's *action policy*. This policy specifies—perhaps probabilistically—what action the agent takes in any given state. Traditionally, reinforcement learning methods specify how the agent's policy evolves as a result of its experience. Crucially, the environment is memoryless: by the titular Markov property, the distribution of next states depends only on the current state and the action taken by the agent in it. History plays no direct role in the dynamics of MDPs.[1]

We may associate to each state its *value* under a given policy, which is the return we expect the agent to reap if it started in this state and followed the policy. Likewise, we may associate to each state-action pair its value under a policy. This is the return we expect the agent to get if it started in this state, took that action, and followed the policy from then on. Although many reinforcement learning algorithms require the agent to maintain an estimate of one of these value functions, this is not strictly required. But even if the agent makes no use of these functions, they remain sensible objects for us theorists to analyze.

The Markov property is of fundamental importance to reinforcement learning, since it grounds the so-called Bellman recurrence. To explain this recurrence, we introduce a bit of notation.

The set of actions is denoted $\mathcal{A}$ and the set of states is denoted $\mathcal{S}$. We think of an episode of interaction with the environment as a sequence $S_0, A_0, R_1, S_1, A_1, \ldots$ of random variables, with the state variables S_i taking values in $\mathcal{S}$, the action variables A_i taking values in $\mathcal{A}$, and the reward variables R_i taking values in the real numbers $\mathbb{R}$. By the Markov property, S_{t+1} and R_{t+1} are independent of the entire history of the episode, conditional on the values of S_t and A_t. In other words, the distribution over histories is entirely determined by the conditional distribution $P(S_{t+1}, R_{t+1}|S_t, A_t)$. The agent's policy is denoted $\pi(a|s) = P(A_t = a|S_t = t)$, and the notation is meant to remind us that the agent's actions depend only on the current state. G_t is the (discounted) sum of the agent's rewards, starting at time t. Finally, we let $v_\pi(s)$ denote the value of state s under policy π and $q_\pi(s, a)$ denote the value of the state-action pair (s, a).

[1] This is also true of the agent, relative to a given policy: policies are also memoryless. But of course, the point is that the agent's policy evolves as a result of its experience, so that the distribution of actions does change over time.

Now, we are in a position to exhibit the Bellman recurrence:

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t|S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots |S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \ldots |S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = t] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma \mathbb{E}[G_{t+1}|S_{t+1} = s']\right] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')]
\end{aligned}$$

Thus, the value of a state is the expected reward from being in this state plus the value of the expected next state. Note that the value function occurs on both sides of the identity. A similar equation can be derived for action values.

Many reinforcement learning algorithms use the Bellman recurrence to learn a policy. As such, the Bellman recurrence is of central importance to reinforcement learning. It is easy to show that there is exactly one function, the true value function, satisfying the recurrence. For any other putative value function, there will be a difference between its estimate of the value of s, on the one hand, and the expected next reward plus the estimated value of the expected next state on the other. This difference is called the *temporal difference error*, and is denoted δ:

$$\delta = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$

This quantity plays an important role in the most distinctive reinforcement learning algorithms, as well as in many applications of the framework to cognitive and neuroscience. Its importance owes to the fact that nudging the estimated value of a state in the direction of the prediction error brings the estimate closer to the truth.

The point of reinforcement learning, however, is not to learn the value of states or actions, but to learn how to act—to learn a policy. If we have a reasonably accurate value function, however, we can improve our policy, by choosing actions that are more likely to lead to high-value states (this is called making our policy more *greedy*). Note that the value of a state depends on the operative policy, while the policy in turn depends on the distribution of value across states. Updating our value estimate may lead us to revise our policy, which may in turn cause us to update our value estimate, and so on. Remarkably, under mild conditions, various versions of this basic strategy converge upon an optimal policy—one that achieves the highest possible expected return in the given the environment.

Among these conditions, the most important is that each state is visited at least once. This requirement induces a dilemma between *exploration* and *exploitation*. An agent exploits when it takes what it estimates to be the best action in a give state. In exploration, the agent foregoes the action with maximal expected return in favor of one with lower expected return. This apparent

deviation from decision-theoretic norms is justified by the fact that the agent's value estimates may be unreliable, being formed on the basis of too little evidence. Only by exploring widely enough can the agent be confident in its picture of the world and in its plans for navigating it.

Computationally, there are many ways to ensure that the agent explores in a sensible way. The most straightforward is to simply have the agent perform a random action some small fraction of the time. It is then advisable to reduce this fraction over time: as the agent's policy converges to the optimal one, it has less and less need to deviate from it in search of alternatives. A more sophisticated way to encourage exploration, and a potentially more psychologically realistic one, is to reward information gain. To do so, the agent tracks its uncertainty about its value estimates, often by simply counting the number of times it has visited—and hence gotten information about—each state or state-action pair. Actions associated with high uncertainty are then deliberately overvalued, and hence more likely to be selected.

These strategies are interesting, in that they reveal that choosing the worse action, or over-estimating the value of a state, can be not only beneficial but also necessary to achieve the distal goal of acting optimally or accurately valuing each state. Exploration takes on a particularly important role in *non-stationary* environments, in which the environmental dynamics evolve over time. In such environment, a perfectly accurate value function at one time may be drastically mistaken a few time steps later—think of a fallen tree blocking what used to be the fastest route home. Frequent and continuing exploration is often more appropriate in such a setting.

3.2 Reinforcement Learning in the Cognitive Sciences

Since the mid-nineties, reinforcement learning has enjoyed considerable explanatory success in the cognitive and neural sciences. In the late eighties and early nineties, the framework was found to elegantly explain and unify an array of results on animal learning that the then-prevalent conditioning paradigm struggled to accommodate. These successes were soon followed by the discovery that the algorithmic structures posited by reinforcement learning had close neural correlates.

[TODO: explain connection between TD learning and classical conditioning]

[TODO: give a sense for the breadth of contemporary psych using RL]

[TODO: overview of role of RL in neuroscience]

4 RL and representation learning

In this section, we cover some of the connections between reinforcement learning and *representation learning*. Representation learning is an interdisciplinary area of research spanning machine learning and cognitive science, focused on elucidating representational structure and the ways such structures can be learned from experience.

Representation learning is particularly relevant to reinforcement learning for two reasons. First, applications of reinforcement learning to real-world scenarios face a scaling challenge. As the size of the state and action spaces increases, the task of finding an optimal, or nearly optimal, policy becomes computationally intractable. This is because in such environments, states are rarely visited more than once, and often not at all. Moreover, when a state is visited, only one out of several possible actions is taken. Thus, even with extensive exploration, the agent cannot sample but a tiny fraction of the problem space, and hence cannot form well-informed value estimates.

Unfortunately, most real-world applications of reinforcement learning, whether natural or artificial, face intractable problem spaces [CITE: Gershman and Daw 2017, others]. Indeed, sensory stimulations are usually continuous and high-dimensional, defining uncountably infinite state-spaces. Likewise, at any given time, the agent faces a vast range of possible actions. Straightforward reinforcement learning algorithms are essentially powerless in the face of this complexity.

Representation learning helps to tame this complexity by learning efficient ways of summarizing information about the problem. In particular, representation learning can help the agent manage the size of the state space, by learning useful groupings of states. It can also help manage the size of the action space, by

5 Philosophical work on RL

Philosophers have paid scant attention to reinforcement learning. In this section, we summarize the extant literature on the subject. Some of the literature focuses on the implications of RL for our understanding of the (human) mind, while some use tools from the philosophy of mind and action to elucidate key concepts in RL, especially as found in machine learning. Finally, some authors place RL in the larger context of decision theory.

5.1 RL and value

In a pair of papers, Julia Hass argues that the success of the reinforcement learning approach in cognitive science holds lessons for our understanding of the mind.

[CITE: Haas 2022] argues that the notions of *reward* and *expected value* that lie at the core of the reinforcement learning framework can be used to analyze the folk-psychological notion of *desire*. To desire something, according to Haas [CITE: Haas in prep], is to subpersonally attribute a subjective reward or expected value to that thing. The notions of reward and expected value Hass uses are supposed to be the *lingua franca* of reinforcement learning. Thus, if Haas's thesis is correct, the reinforcement learning framework offers a powerful tool for understanding a psychological state of perennial philosophical interest.

Haas also draws on the success of reinforcement learning models of selection to argue that evaluation is fundamental to the mind. Selection encompasses a broad range of cognitive tasks in which the mind must select from among a set of alternatives. For example, in visual perception, the mind must decide where—and on what features—to focus its attention. In action-selection, the mind must select one among a set of alternative actions. [CITE: studies] And so on.

Many such selection tasks have been analyzed through the lens of reinforcement learning. In such models, the reinforcement learning agent must choose among several options. In many cases, the agent learns to maximize expected return by maintaining a representation of each option's value (perhaps relative to a given state). Options are then selected on the basis of their estimated values.

Assigning value to options is therefore a core component of many successful models of various core mental processes. In light of the explanatory success of such models, Haas argues that a wide range of basic cognitive processes implicate valuation.

5.2 RL and curiosity

[Explain Nagel's work on RL and curiosity]

5.3 RL and action

In a pair of papers, Butlin has argued that reinforcement learning systems are agents. Butlin draws on [CITE: Dretske]'s theory of action to develop an account of *minimal action*. On this minimal conception, behavior must satisfy two conditions to be an action: first, it must be selectively produced in response to certain environmental conditions, where this selectivity is the result of a learning process. Second, the learning process must function to produce behavior that is instrumentally valuable. That is, the instrumental value of the behavior must explain why it was learned, and this explanation must "go through" the learning process.

This conception of action is minimal in that it does not require any of the sophisticated agential capacities that (sometimes) accompany human action, such as deliberation, planning, coordination, and so on. As such, the behavior of primitive organisms and artificial machines is candidate for minimal action. Still, since behavior can only be instrumentally valuable relative to a goal, this minimal conception requires actions to be goal-directed. Indeed, Butlin also characterizes actions as activities subject to norms, where the relevant norms derive from goals rather than functions (the two characterizations are supposed to be coextensive). Minimal actions are thereby set apart from merely functional behavior, such as the beating of a heart. The latter has a function—to pump blood—but serves no goal.

Of course, the distinction between functions and goals invites explication. Butlin adopts the *selected-effects* theory of biological function [CITE: Garson 2016]. This theory maintains that a component of an organism functions to perform some activity just in case that component was selected to perform that activity. A component is selected to perform an activity just in case its performance of that activity explains its stable, continued existence. Thus, the function of a component is to bring about behavior that has previously caused this component to be stabilized.

By contrast with functions, which are set by processes of selection, goals are set by learning processes.

> Minimal agency requires learning to produce outputs selectively for their contributions to good performance over an episode of interaction with the environment.
>
> ...
>
> When systems undergo learning which is sensitive to the contributions of outputs to performance over episodes, and promotes performance of a particular kind, they come to pursue the goal of good performance through their outputs, and their activity can be evaluated according to whether it is conducive to this goal.
>
> ...
>
> They have goals as opposed to functions.
>
> [CITE Butlin 2024: 27]

It is not immediately clear how this characterization distinguishes functions from goals. For example, the heart also produces outputs (i.e., beats) for their contribution to good performance (i.e. survival) over an episode of interaction with the environment (i.e. a lifetime). Yet Butlin insists that hearts merely function to pump blood, and do not have pumping blood as a goal. The main difference seems to be that the heart's beating is the product of a history of selection, but does not involve learning. Not any kind of learning will ground goals, however:

> For a biological or artificial system to satisfy this account of agency the way in which it learns must allow information about subsequent performance to influence the probability that an output will be repeated, under environmental conditions of a given kind.
>
> [CITE Butlin 2024: 28]

Thus, the learning process must be sensitive to a given output's contribution to subsequent performance: learned behavior must be learned because of its contribution to future performance. In summary, agency requires behavior to be the product of a learning process sensitive to instrumental value.

I note in passing that some cases of behavior that seem clearly agential, such as a chick's pecking behavior, are plausibly not learned but instead innate ("grown," to use Chomsky's phrase). Other cases of minimal agency, such as paramecia's locomotion, are even more clearly not the product of a learning process (though their status as agents is perhaps negotiable). Butlin's account, however, wrongly classifies such behavior as non-agential. It is also unclear how Butlin's account fares in cases of one-off actions. The behavior of someone throwing a dart at a dartboard for the first time is not the product of a learning process (at least, not in any direct sense), yet it is clearly agential. Admittedly, this is not a case of minimal action. But if learning is not required for more sophisticated kinds of actions, why should it be required for minimal actions? [Perhaps because in the case of high-level actions, goals can be set by intentions, whereas no such intentions are available for minimal actions, in which case a history of learning has to step in?]

Setting aside these concerns about Butlin's notion of minimal agence, we turn to his contention that reinforcement learning systems are agents. Recall that the first condition on minimal agency is that the system must learn to selectively produce behavior. Barring concerns about whether artificial systems can learn anything at all (concerns I do not share), reinforcement learners clearly meet this condition. The second condition requires this learning process to be sensitive to instrumental value. Again, it is reasonably clear that this holds of reinforcement learners. Almost all reinforcement learning algorithms function to increase the probability that actions with high expected value are selected. That is, they do not merely happen to increase the probability of high-value actions. Rather, they increase the probability of certain actions because they have a high expected value. Actions with high expected value are expected to contribute most to the long term goal of maximizing cumulative reward. Thus, the learning

algorithm of a reinforcement learner selectively reinforces actions according to their expected contribution to long-term performance: they are sensitive to instrumental value. Thus, reinforcement learners are minimal agents.

Butlin's account has a number of interesting philosophical consequences. First, it implies that supervised learning systems, in contrast to reinforcement learning systems, are not agents. Supervised learning algorithms learn by comparing their output on a given input to the correct output. If their output is incorrect, their internal state is modified so as to increase the probability of producing the correct output. Crucially, the feedback given to a supervised learner does not include the identity of the next input (as it does in reinforcement learning). Thus, a supervised learner cannot learn to produce a certain output because of its influence on the downstream sequence of inputs it will receive (indeed, its outputs typically *have* no such influence). As a result, the notion of instrumental value does not apply to supervised learner. A fortiori, their learning algorithm cannot be sensitive to the instrumental value of their outputs, and hence they are not agents.

Here I register a skeptical comment. It is possible to view any supervised learning problem as a special case of a reinforcement learning problem. Let me illustrate this with a binary classification problem for simplicity (the point applies to all forms of supervised learning). In such a task, the learner's task is to learn a target function $f : \mathcal{X} \to \{0, 1\}$ that provides the true classification for each instance $\boldsymbol{x} \in \mathcal{X}$. To learn this function, the learner has access to some training data $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ consisting of pairs of samples $\boldsymbol{x}_i$ along with their correct classification $y_i = f(\boldsymbol{x}_i)$. The goal of the supervised learner is to achieve the best performance possible on this training data. There are various ways of measuring performance; one popular way is to simply count the number of training examples which the learner misclassifies. This supervised learning problem can be construed as a reinforcement learning problem as follows. Each sample $\boldsymbol{x}_i$ corresponds to a state in the reinforcement learner's MDP. The agent's actions, in each state, are to output either 0 or 1. The reward for each action is given by $f(\boldsymbol{x}_i)$. After each action, the state updates to $\boldsymbol{x}_{i+1}$ (or the episode ends, or wraps back around to x_1, if $i = N$). Clearly, maximizing expected reward over the course of an episode is equivalent to minimizing the number of misclassified states. Thus, the supervised learning problem is also a reinforcement learning problem.

Of course, there is little point in modeling a supervised learning problem in this way. MDP are particularly suited to model situations in which the agent's actions have an effect on the environment. But precisely this feature is lacking from this example: both available actions in each state lead to the same next state. Nonetheless, it does not seem out of place to talk of a supervised learning system having as its goal correct classification, and the reinforcement learning framework seems to support this attribution through such hackneyed cases. Such examples therefore put pressure on Butlin's contention that supervised learners are not agents.

The existence of active supervised learners exerts additional pressure on Butlin's thesis. Active learning is a sub-field of supervised learning in which

the data is not presented to the learner in a default, or random sequence, as is done in "vanilla" supervised learning. Instead, the supervised learner "decides" which training example to look at next. This decision is often driven by information-theoretic considerations: the model asks about examples which it is most uncertain about [CITE: Gureckis, Murphy]. As such, active learning models interact with their environment in a way that impacts the distribution of examples they see. They therefore seem to meet Butlin's criteria for agency. Yet is it not clear that they are agents.

Butlin also uses his framework to argue that model-based reinforcement learning agents are capable of acting for reasons. Butlin adapts Mantel's [CITE: Mantel] account of acting for reasons. On Butlin's construal, to act for reasons is to select actions via a general-purpose capacity to select actions represented as conducive to one's goals. Model-based reinforcement learners maintain a representation of the environmental dynamics: they come equipped with, or learn, the environment's transition function.[2] When choosing actions, they use their representation of the transition function to compute (an approximation to) the action with the highest value. Typically, this is the action they choose to perform. According to Butlin, this means that "they act on representations of facts about transitions in ways which are, in their fundamentals, the same as the way in which humans act on instrumental beliefs" [CITE Butlin 2024: 32].

Again, it is not clear to me why model-free reinforcement learners fail to meet this condition. That an action has a higher expected value than all alternatives seems like a consideration in favor of performing it, and it would seem that it is precisely on the basis of such considerations that model-free reinforcement learners choose actions. In other words, they seem to engage in reasoning about which available action is most conducive to their goals. Of course, this reasoning is much simpler than the kinds of computations that model-based learners perform: it only involves a comparison operation, whereas model-based decision-making involves tracing the consequences of an action a few steps into the future. But Butlin (rightly) does not consider the complexity of the underlying reasoning to make the difference between acting for reasons and (merely) acting.

5.4 Rational RL

[CITE: Huttegger 2017] places reinforcement learning in the context of Bayesian decision theory. Hutteger's presentation of reinforcement learning abstracts from almost all algorithmic details, characterizing instead the underlying structure that such algorithms induce on behavior. In addition, Hutteger appears to focus on a specific form of reinforcement learning, so-called policy learning (which he calls the "basic model of reinforcement learning"). Most reinforce-

[2]There are in fact two kinds of model-based reinforcement learners. Sampling model-based learners can sample from the transition function, while distribution model-based learners can compute this function. The latter enables much more sophisticated planning than the former, and typically has greater computational complexity. Presumaly, Butlin's discussion refers to distribution model-based learners.

ment learning algorithms learn a good policy by learning accurate state or action values, from which the optimal policy follows (almost) trivially.[3] Policy learners eschew value learning. Instead, they assign a score to each possible action in a given state and choose actions with probability proportional to their score.[4] Action scores are updated so as to improve the corresponding policy. Importantly, an action's score need *not* be that action's value.

Hutteger's main results follow the playbook of *ecological rationality* [CITE: Gigerenzer et al.]. This school of rational analysis seeks to uncover the conditions under which behavior that might appear, in the abstract, less than ideally rational, is in fact optimal. While these conditions are sometimes taken to lie beyond the agent (in the environment), Hutteger's focus is on internal consistency. Thus, he addresses the question, what must an agent believe about his situation for the basic model of reinforcement learning to be an optimally rational way to navigate the world? Optimal rationality here is given by Bayesian decision theory.

[Note: I'll explain Huttegger's answer somewhat carelessly, since I don't know how to express it more carefully]

Huttegger's answer is that the basic model of reinforcement learning is optimally rational when the agent's credences satisfy a kind of symmetry called *partial exchangeability*.

...

5.5 RL and resource rationality

[Explain Icard's take on RL]

[3] The optimal policy follows trivially from the action values: simply pick the action with the highest value. To compute the optimal policy from state values, the learner needs a (distribution) model of its environment, since it needs to compute the expected value of taking each action in a given state, and doing so requires knowledge of the likely consequences of each action.

[4] [CITE: Sutton and Barto] call this score the *eligibility vector*.