Reinforcement Learning and Artificial Agency

Patrick Butlin

Forthcoming in *Mind & Language*; please cite published version

Abstract

There is an apparent connection between reinforcement learning and agency. In computer science entities controlled by reinforcement learning algorithms are standardly referred to as agents, and the mainstream view in the psychology and neuroscience of agency is that humans and other animals are reinforcement learners. This paper examines this connection, focusing on artificial reinforcement learning systems and working from the assumption that there are a range of forms of agency. Artificial reinforcement learning systems satisfy plausible conditions for minimal agency, and those which use models of the environment to perform forward search are capable of a form of agency which may reasonably be called action for reasons.

1. Introduction

Reinforcement learning (RL) is a branch of machine learning with an apparently intrinsic connection with agency. In their seminal textbook on the subject, Sutton and Barto (2018, p. 1) define reinforcement learning as, 'learning what to do – how to map situations to actions – so as to maximize a numerical reward signal'. In the standard reinforcement learning framework the entity that is controlled by the RL algorithm, which moves through the real or virtual environment, is always referred to as an agent. This piece of terminology is only very weak evidence that RL systems really are agents, but it does prompt a philosophical question: what does RL have to do with agency? In this paper my aim is to make progress on this question.

There are two possible directions from which this topic could be approached. One would focus on artificial intelligence, and ask: in what respects are artificial RL systems agent-like, or what aspects of agency do they display – if any – in virtue of their use of RL? This version would be timely because the DeepMind systems which have achieved superhuman performance in Go, chess, shogi and various computer games all worked by combining RL algorithms with multi-layer artificial neural networks (Mnih et al. 2015, Silver et al. 2016, Silver et al. 2018, Schrittwieser et al. 2020). Deep RL systems have also begun to be used for some practical applications (Whittlestone et al. 2021). Alternatively, since the RL framework has become the dominant paradigm in the psychology and neuroscience of evaluative cognition and action selection (Schultz et al. 1997, Niv 2009, Dolan & Dayan 2013, Gershman & Daw 2017), one could ask

about the significance of RL for human and animal agency. Here the question would be: what does the apparent fact that humans and other animals engage in RL have to do with their agency? These two questions are entwined, but in this paper I will focus on the first. I will use ideas which have been developed in the context of theorising about human and animal agency, but AI will be my main target.

Understanding agency in AI is important for several reasons. One reason is that goal-directed AI agents are more likely to exhibit unpredictable or power-seeking behaviour, which may make powerful systems especially dangerous (Omohundro 2008, Bostrom 2014). This thought has prompted interest in the nature of agency among AI safety researchers (e.g. Kenton et al. 2022). A second reason is that if artificial systems can be agents we will at some point confront problems of moral agency and responsibility in AI (Wallach & Allen 2008). A third is that agency may matter for moral standing in AI (Kagan 2019, Shulman & Bostrom 2021). However, I will focus directly on the relationship between reinforcement learning and agency, rather than addressing any of these topics.

According to a traditional view in philosophy, which remains influential, there is an important form of agency which is distinctively human, sometimes called action for reasons. Accounts of this form of agency emphasise self-awareness, the ability to select principles to be followed in action, and the ability to conceive of reasons as such. Jamieson (2018) associates such accounts with the view that there is a discontinuity between human and non-human animal agency, and they certainly suggest that there are demanding requirements for paradigmatic agency. However, there is more to agency than just this form. Philosophers have also explored concepts of minimal and primitive agency (Barandiaran et al. 2009, Burge 2009), and forms of agency which humans share with other animals (Glock 2009, Sebo 2017, Jamieson 2018). Proponents of demanding accounts of action for reasons acknowledge that related forms of agency exist in children and other animals (Velleman 2000, Schlosser 2012). Korsgaard (2018) states a position of this kind in arguing that animals are agents because they engage in 'representation-governed locomotion', but humans are distinctively rational agents because we reflect on our motives and are aware of our actions as our own. It may be that there are levels of agency – perhaps overlapping and blending into one another – which depend on different levels of cognitive sophistication, such as those discussed by Dennett (1996), Bratman (2000) and Papineau (2001). If this is right, then forms of RL may play a role in distinguishing levels of agency.

From the perspective that there are a variety of forms of agency, perhaps forming a hierarchy, it becomes possible to view apparently conflicting positions in the philosophy of action as accounts of conditions for different forms. For example, Velleman (2000) argues for a

demanding account of 'autonomous action', over an alternative belief-desire account, but also acknowledges that the belief-desire account may capture a phenomenon he calls 'purposeful activity'. This matters for my purposes because I will argue that two forms of RL are respectively sufficient for two forms of agency, but I do not deny that there may be further important forms of agency for which the conditions are more demanding. This is not to say, however, that all accounts of forms of agency are equally valid. Some will be more successful than others in identifying joints in nature or phenomena which matter for further theoretical or practical purposes. I will also shortly argue that some accounts of minimal agency are too weak to delineate forms of agency at all, with the phenomena they describe perhaps better seen as preconditions for agency.

Minimal agency is the first topic I will discuss, in sections 2 and 3. I will argue that it is possible for artificial systems to be agents with goals, in opposition to the view that only biological self-maintenance can give rise to the normativity which is necessary for agency. I will describe the 'model-free' form of RL and argue that it is sufficient for minimal agency. Then in sections 4 and 5 I will describe 'model-based' RL and argue that it is sufficient for action for reasons. The difference between model-free and model-based RL is that in the latter the system uses a transition model of its environment, which is a representation of the probabilities that actions will lead to particular outcomes when taken in given initial states, to select actions. This means that it represents facts which count in favour of or against possible actions and selects actions on the basis of these representations.

2. Minimal agency

I begin by discussing accounts of minimal agency. In this section I first criticise an account of minimal agency by Barandiaran et al. (2009) which makes it a primarily biological phenomenon, then present an alternative proposal by Dretske (1985, 1988, 1993, 1999). Dretske's theory is more readily compatible with agency in artificial systems, and also does a better job than the bio-agentic account of grounding the normativity of agency. However, Dretske's theory is too weak – it attributes agency to systems which do not pursue goals through interaction with their environments – so I propose a revision to fix this problem.

To be an agent, a system must engage in goal-directed interaction with an environment. This means that its interaction with the environment must be governed a norm, at least in the weak sense of a non-arbitrary standard of success or correctness. A system with a goal is subject to a norm because it can perform better or worse by producing outputs which are more or less conducive to that goal. However, having a goal is a specific form of norm-governance. Biological

sub-systems such as the heart and artifacts such as mousetraps have functions, so their activities are subject to norms, but they do not seem to be agents or to pursue goals. So two crucial tasks for a theory of minimal agency are explaining the sense in which the activities of minimal agents are governed by norms, and showing that they have goals rather than functions.

Normativity in agency is a central concern of theories of *bio-agency*, which Meincke (2018, p. 65) defines as 'the intrinsically normative adaptive behaviour of … organisms, arising from their biological autonomy.' In an example of a bio-agentic account, Barandiaran, Di Paolo and Rohde (2009) argue for three necessary and jointly sufficient conditions for 'genuine' agency, which they call *individuality*, *asymmetry* and *normativity*. The individuality condition is that there 'must be a distinction between the system and its environment' (p. 369), and Barandiaran et al. argue that this is not trivial, because in the case of artifacts the distinction between an object and its environment is drawn by observers, for their own convenience. Agents must 'define' their own individuality, which all living organisms do in virtue of their self-generating, self-maintaining character. The asymmetry condition is that an agent 'systematically and repeatedly modulates its structural coupling with the environment' (p. 372), meaning that the agent's internal state must cause changes in the nature of the agent-environment interaction. The normativity condition is that an agent's interaction with the environment must be regulated so as to meet some norm, and Barandiaran et al. insist that the norms in question must be determined by the agent's 'own viability conditions' (p. 376). These conditions add up to an account of agency which is very liberal in some respects; Barandiaran et al. write that a bacterium performing chemotaxis is an agent, and plants also seem to satisfy their conditions.

For Barandiaran et al., the normativity required for agency can only arise from self-maintenance of the kind found in living organisms. They do not claim that this entails that artificial agency is impossible, but they do write that 'systems that *only* satisfy constraints or norms imposed from outside should not be treated as models of agency' and that agents' actions must contribute to their self-maintenance (p. 381). Artificial RL systems typically optimise reward functions which are externally imposed and unrelated to their persistence, so their account is incompatible with agency in these systems.

There are two problems with this view. First, there are many cases in which living organisms perform actions which do not promote their own self-maintenance, but instead contribute to reproduction or the fitness of their kin or group. The imperative for organisms to reproduce is externally imposed in that it is genetically inherited, yet behaviour taken towards this end is just as clearly agentic as towards any other (for one thing, it can call upon the full range of cognitive capacities associated with agency). So self-maintenance is not the only source of norms by which

the activities of living organisms are regulated. Second, the orthodox view in philosophy of biology is that traits of organisms can have functions in virtue of the effects for which they were selected (Garson 2019). This means that a compelling alternative account of the grounds of norms in biology is available. Having a biological function is not sufficient to be an agent, as the example of the heart illustrates, but if the normativity of function can be explained in terms of selective history – as opposed to self-maintenance – it is possible that this is also true of the normativity involved in agency.

Dretske's theory, in contrast, grounds normativity in etiology – although in learning, not selection – and is compatible with artificial agency. His proposal is that, for some behaviour *b* of type *B* to be an action, it must be the case that:

- *b* is caused by an internal state *r* of type *R*;
- states of type *R* carry information about a feature of the environment *E*;
- and the system has learnt to produce *B*-behaviours when in *R*-states partly in virtue of the fact that *R*-states carry information about *E*.[1]

This claim is helpfully illustrated by an example from Dretske (1999). Suppose that a bird refrains from eating a viceroy butterfly which lands near its perch, because it has previously eaten a similar-looking monarch butterfly which tasted bad. This behaviour would count as an action, because it would (presumably) be caused by a state which the bird's brain enters when it encounters things with the appearance of monarch butterflies. This internal state would cause the behaviour partly because it has this correlation with the environment – it is because there was an actual monarch butterfly present on the earlier occasion that the bird experienced the bad taste, and therefore that a connection was established between the internal state and the form of behaviour it exhibits. This theory is less complex than it appears: it is just the claim that action is behaviour that the system has learnt to produce selectively in particular circumstances, presented in a way which reflects Dretske's further concern with naturalising representation.

According to Dretske, when these conditions are met states of type *R* are used as indicators of *E*. To say that they are used in this way is to go beyond the claim that they carry information about *E*. It means that these internal states are representations with correctness conditions; in the butterfly case the internal state misrepresents, because it is being used as an indicator of monarch butterflies, but is triggered by a viceroy. Dretske thinks of action as behaviour governed by thought, and takes this to mean that actions must be caused by internal

[1] This theory is described in Dretske (1999), which focuses specifically on what distinguishes agents from non-agents. In *Explaining Behavior* (1988), where he aims to account for the explanatory role of reasons, Dretske also describes how internal states playing a desire-like role can contribute to behaviour. So a more demanding account could be extracted from this work.

representations in virtue of their content. His theory is supposed to capture this idea because it requires that correlations between representations and the environment explain why those representations cause behaviours. In his terms, content is a 'structuring cause' of behaviour.[2]

Dretske's theory takes an attractive approach to grounding normativity and entails that some artificial systems are agents. However, it is too liberal because it does not capture the goal-directed aspect of agency, focusing instead on the idea of actions as caused by representations in virtue of their content. I will now explain these three points in turn.

Dretske's approach to grounding normativity is a version of that developed by teleosemantic theorists, such as Millikan (1984). The core idea is that a feature of a system can come to have a function by being selectively retained or reproduced as a result of some activity that it performs or effect that it produces. On this view whether an entity has a function depends on whether its existence is explained, etiologically, by what it can do. This is an account of normativity (of the kind which is relevant here) because any entity with a function is subject to a standard of success or correctness derived from that function. The most familiar version of this etiological approach claims that natural selection gives rise to functions of biological traits which are selected for fitness-promoting effects, and this theory has been widely discussed, defended and refined (e.g. Garson 2019). Mossio et al. (2009) proposed a theory of biological function based on self-maintenance, which Barandiaran et al. cite in support of their view, but Artiga and Martínez (2016) argue that this approach has no advantage over the etiological one.

Dretske's view is that agency requires functions to be established by learning, as opposed to evolution. However, like evolution, learning is a process in which features of systems are selectively established and retained for their effects. In biological cases learning is responsible for many of the traits and behaviours that adapt organisms to their environments. The idea that learning can ground functions has recently been defended in detail by Shea (2018), who describes it as a process by which complex systems develop so as to produce outcomes robustly.

Turning to the application of his theory to artificial systems, it is striking that Dretske argues in a series of papers (1985, 1993, 1999) that there can be no such thing as genuine artificial intelligence. This is because, in Dretske's view, intelligence is agency and agency requires learning. Artifacts can easily be constructed which produce outputs under certain environmental conditions, in virtue of correlations between these conditions and internal states, but Dretske's view is that in such cases the correlations do not explain the connections between internal states

[2] Hofmann and Schulte (2014) argue that Dretske's theory does not succeed in explaining how the content of mental states can cause behaviour. But this is compatible with its providing an attractive account of a minimal form of agency.

and outputs in the right way. Artifacts' dispositions are explained by the actions of their designers, not by their sensitivity to the environment. What this argument neglects is that learning is possible in artificial systems. Like learning in animals, machine learning involves endogenous change in response to feedback, which proceeds according to an externally-imposed learning algorithm and tends to improve performance by some standard. Human engineers exercise some control over this process, in some cases, by providing the inputs and feedback by which machine learning systems are trained. But humans also train animals, and this in no way entails that the animals are not learning.

In fact, Dretske's theory entails that some machine learning systems are agents. To illustrate this claim I will discuss a system which is *not* an agent; this example will then also show the flaw in Dretske's scheme, and motivate my proposal.

AlexNet (Krizhevsky et al. 2012) is an image classification system based on a deep convolutional neural network, trained by supervised learning. Its training worked in the following way: at each step it was given as input an image sampled from a labeled corpus of over one million; this input caused activation to flow through the network, leading to an output assigning probabilities to each of 1000 categories; then the correct label for the image was provided as feedback, with the network weights adjusted depending on the difference between the output and this feedback. This regime led AlexNet to take a form in which it would reliably produce outputs assigning the highest probability to the correct label in response to input images from a held-out portion of the corpus.

AlexNet satisfies Dretske's conditions because, after it has been trained, patterns of activation in the network cause outputs of particular types. These patterns of activation are also correlated with input types, and it is because the patterns are correlated with particular input types that, as a result of learning, they cause particular outputs. For example, learning will cause patterns of activation which are highly correlated with images of crowns to become causally linked to 'crown' outputs. Provided that we think of input images as forming part of its environment, AlexNet therefore learns to produce outputs selectively in response to environmental features.

This is a problem for Dretske's theory because AlexNet does not seem to be an agent. It does acquire a function through its training – its function is to classify images – but it does not seem to pursue a goal. One telling point is that the inputs AlexNet receives in training are probabilistically independent of each other and of its outputs. In particular, AlexNet's outputs do not affect its subsequent inputs. This means that AlexNet can only ever learn to produce the right output for each individual input, as opposed to learning how to pursue goals through interaction with its environment. The latter involves producing outputs because they will affect

subsequent inputs, making it possible to achieve goals through episodes of interaction. Dretske's theory goes wrong because producing outputs in a representation-guided way is compatible with the form of behaviour exhibited by AlexNet as well as that which is characteristic of agency.

I therefore propose that minimal agency requires learning to produce outputs selectively for their contributions to good performance over an episode of interaction with the environment. One way for an output to contribute to good performance is for it to be one step in a series by which some goal can be achieved. 'Good performance' here just means the kind of performance which tends to be selected for by the learning process. 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. So systems of this kind are subject to a specific kind of norm, which is distinctive of agency. They have goals as opposed to functions. I take this to be a way of developing the idea of goal-directed interaction with which I began this section.

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. As I will explain, this can be done in various ways. However, the way in which AlexNet learns does not meet this description, because the feedback to AlexNet's outputs only includes information about the correct response to the previous input. Updates concerning a given output are then completed before the next input is provided. In contrast, agents must learn in a way which is sensitive (at least indirectly) to input-output-input contingencies.

Another way to put the idea that agents pursue goals is to say that they select outputs in a way which depends on the instrumental value of these outputs. Following Dretske, I take learning to be crucial to explaining how agents select actions, so my account of agency can be satisfied by a system which learns in a way which is sensitive to instrumental value. As I will explain in the next section, model-free RL systems fit this brief precisely.

3. Model-free reinforcement learning and minimal agency

Model-free reinforcement learning systems satisfy the conditions for minimal agency just proposed. In this section I give a brief introduction to reinforcement learning,[3] focusing on aspects which are relevant to my arguments, then explain why this is so.

[3] For a more detailed introduction for philosophers, see Haas (2022). The canonical textbook in the field is Sutton & Barto (2018).

RL algorithms are methods by which systems can learn, from interaction with an environment, how to behave in that environment so as to achieve an objective. To aid the development of such algorithms researchers have adopted a standard way of modelling environments, as Markov decision processes (MDPs). MDPs have the following elements: a certain number, perhaps infinite, of possible *states*; a range of *actions* available to the system in each state; and a numerical range of possible levels of *reward* that the agent can receive.[4] Time is modelled as passing in discrete steps, and at each time-step the state that the system enters and the reward that it receives depend solely but perhaps probabilistically on the previous state and the action selected. A *transition function* describes which new state the system will enter after performing each action, in each initial state, and a *reward function* describes how much reward it will receive.[5] The system's objective is to maximise the reward function, so there is a sense in which the objective is built into the environment.

More precisely, the system's objective is to maximise the amount of reward it receives over the entire episode of its interaction with the environment. This means that at each time-step, the immediate consequences of an action are less important than the long-run cumulative reward which can be expected to follow from it, which is called *value*. The value of an action can only be defined relative to assumptions about how the system will behave in subsequent states. Similarly, the value of a state can be defined as the expected long-run cumulative reward subsequent to being in that state, again relative to assumptions about future behaviour. A system's *policy* is a function describing how it will behave in each possible state; actions and states have values relative to policies, and the aim of RL can be described as finding an optimal policy, where this means one that maximises cumulative reward.

The terms which RL researchers use – especially the talk of actions and objectives, but to some extent also reward and value – are problematic for my purpose, because whether RL systems perform actions is the issue at hand. As I have mentioned, it is also standard to call these systems 'agents'. I will use 'outputs' and 'inputs' rather than 'actions' and 'states' where it is worthwhile to be scrupulous, but continue to use the latter terms at some points where I expect this to aid understanding. We should also bear in mind that I have not yet established that RL systems pursue objectives; what is known is that they are designed (more or less successfully) to

[4] RL algorithms have also been developed for so-called 'multi-armed bandit' problems, in which the state does not change. Systems capable of solving these problems need not be agents, so I leave this branch of RL aside to focus on that concerned with MDPs.

[5] Researchers vary in whether they take the domain of the reward function to be states (in which case reward depends on which state the agent reaches when it performs an action) or state-action pairs (in which case reward depends on the initial state and the action selected). I adopt the former convention here.

modify their own input-output dispositions in a way that tends to increase the amount of reward they receive.

I have also mentioned that there is a distinction between model-free and model-based RL. The difference is that model-based methods involve learning a representation, or model, of the transition function. Humans are thought to use both forms (Kool et al. 2018). I will discuss model-based algorithms in the next section; here I focus on the model-free form.

In model-free RL, algorithms are designed to allow systems to learn action values. This makes action selection simple, because a system will maximise the reward function if it always chooses the action with the highest value, relative to the optimal policy, in the current state. In the course of learning the system lacks access to the optimal policy (or the action values for this policy, which entail it), but model-free RL systems use methods for 'bootstrapping' towards this ideal situation. These involve the system taking the actions which it estimates have the highest value, relative to its current estimate of the optimal policy, and gradually improving these estimates based on the reward it receives. For example, in an method called *Q-learning* the system starts with a random assignment of values to actions, and selects actions by either choosing the one with the highest value in the current state, or taking an action at random (which is useful to promote exploration). This causes it to enter a new state and receive a reward, at which point it calculates the difference between the previously estimated value of the action just taken, and the sum of the reward and the estimated value of the best action from the new state (this is called the *temporal difference error*). That is, it calculates:

$$R + \gamma Q(S', A') - Q(S, A)$$

where $Q(S, A)$ is the estimated value of the action just taken in the previous state, $Q(S', A')$ is the estimated value of the best action in the new state, R is the reward just received, and γ is a discount factor. The system updates its estimate for $Q(S, A)$ in the direction of the temporal difference error. Over time this process causes the system's value estimates to become more influenced by its experience of reward in the environment, and hence to converge towards the true values for the optimal policy.

According to my account, systems which use model-free RL will be agents if they learn to produce outputs selectively for their contribution to good performance over episodes of interaction. That this is the case can already be seen from the nature of the RL task and the fact that model-free systems can perform it successfully. The RL task is one in which good performance is defined over episodes, and the feedback to systems does not inform them of the

correct output for the input they have just received. Maximising immediate reward in response to each input does not amount to optimal performance; instead, which output it is best to produce in response to a given input depends on which further inputs will follow, and whether these make it possible to access the greatest quantities of reward. The fact that model-free RL systems can perform this task shows that they learn in a way which is sensitive to instrumental value. Their dispositions to produce outputs in response to inputs are derived from a process which gathers and employs information about subsequent rewards.

How this works in the case of model-free RL is that the update rule is based on the temporal difference error, which includes a term for the estimated value of the best action in the new state (which is observed at the same time as the reward, before the update takes place). This makes it possible for the system to learn to produce an output not because it leads to high immediate reward, but because it leads to a state from which high rewards are expected, based on past experience. The learning process passes information about reward and its accessibility back along chains of possible actions, requiring only that these actions are tried enough times.

Dretske's conditions for agency are implicit in my account, and we can see that they are also satisfied. Model-free RL systems develop input-output dispositions through learning that depend on correlations between inputs as they represent them and states of the environment. These correlations are taken for granted in much RL research (the environment is assumed to be 'wholly observable'), but it is crucial for RL that systems' internal states are correlated with the features of the environment that determine rewards and subsequent observations.

Because model-free RL systems satisfy my account of agency, they each have and pursue a goal – specifically, the goal of maximising cumulative reward. What this means is that their input-output dispositions are formed through a process which tends to bring about a certain kind of result (maximum reward) from episodes of interaction with the environment, and can be explained by the contributions they make to bringing this about. These facts ground norms on outputs: outputs are better if they make greater contributions to that kind of result (in roughly the same sense in which organs or artifacts can perform better or worse). This is the basic principle of goal-directed agency that an agent should prefer actions which are more conducive to its goal. My argument is that if an entity learns to produce outputs for their instrumental value, then that entity must be an agent pursuing a goal.[6]

4. Model-based reinforcement learning and action for reasons

[6] For discussion of some further aspects of my account of minimal agency, see Butlin (2022).

In this section I turn to model-based RL. I begin by describing this form of RL and showing that systems using it are agents by my account. I then give an initial argument for the claim that model-based RL agents act for reasons, which I develop further in section 5.

Model-based RL systems are those that learn the transition function. That is, they learn which new states are likely to follow from their actions, given the states in which those actions are taken. There are different ways in which transition models can be used to support the selection of rewarding actions, but one way is to combine the transition model with either a representation of the reward function, or a representation of the state value function, to calculate how much reward can be expected from possible sequences of actions starting from the current state. This process is called *forward search*. Model-based RL with forward search is used in AlphaGo (Silver et al. 2016), AlphaZero (Silver et al. 2018), and MuZero (Schrittwieser et al. 2020).[7] These systems use neural networks to model transitions and learn value functions, but in simpler environments it is also possible to do this using lookup tables. From this point on I will use 'model-based RL' to refer only to systems which also use forward search.[8]

In contrast to model-free RL, the model-based form does not involve making updates after each output which directly affect the likelihood of that output's being repeated. Model-based RL does not involve storing action values or a representation of a policy which determines which action will be selected on a given occasion. Instead, at each time-step observations of the environment are used to update the model, and actions are subsequently selected by a process of inference involving the model. This means that to explain why a model-based system has produced an output it is necessary to appeal to the process of inference or reasoning that immediately precedes it, as well as learning.

In model-based RL this combined process – of learning and reasoning – is sensitive to instrumental value and designed to promote good performance over episodes of interaction with the environment. The key point is simply that in forward search model-based RL systems look more than one step ahead. Rather than selecting the actions that yield the most immediate reward, they select those that promise the greatest cumulative reward over a longer period. If they learn state values rather than the reward function, their sensitivity to future reward extends beyond the furthest reach of their forward search. Model-free RL works by gradually passing

[7] Halina (2021) gives more detail on how forward search is used in AlphaGo, in the context of a discussion of the system's creativity. She argues that forward search (in this case, Monte Carlo tree search) constitutes planning or 'mental scenario building'.

[8] Thus excluding algorithms in the Dyna family (Sutton 1991), which learn a model of the environment but use it only to generate 'simulated experiences', which help to refine the action value function, which is in turn used to select actions.

information about reward backwards along sequences of possible actions, but model-based RL looks forward along such sequences to forecast future reward.

This means that, if my arguments so far have been successful, model-based RL systems are agents and pursue the goal of maximising reward. I now want to argue that they act for reasons, because they represent facts that count in favour of their actions, given their goals, and they have a general-purpose capacity to select actions which are conducive to their goals, given the facts as they represent them.

The facts in question are those described by the transition function. These are facts of the form: from state s_1, if the agent performs action a, the probability that the next state will be s_2 is x. A fact like this might count in favour of action a in s_1 because s_2 is a state from which a large quantity of reward is accessible and x is high. Whether this means that action a should be performed, however, depends on what other routes to reward are available from the initial state. These facts play the same role as those that are thought of as reasons for action in familiar cases. For example, the fact that it is likely to improve the taste (for me, given what I like) is a reason for me to add pepper to my soup. It does not follow that I have most reason to do this, though, because I may have stronger reasons to take other actions: it may be that chilli oil would be equally likely to cause a greater improvement. Facts about transitions do not count in favour of certain actions intrinsically, but only relative to the agent's goals. A different agent could operate in an environment with the same transition function but a different reward function. But this is again as in the familiar case of reasons – the facts which are reasons for me to season my soup have their status as such only in virtue of the particular goals which I pursue. Model-based RL systems are defined by their capacity to learn and represent facts about transitions.

Furthermore, model-based RL systems select actions in ways which respond to the extent to which those actions are favoured by the facts as they represent them. Forward search involves using representations of facts about transitions to calculate the expected returns from possible sequences of actions, and in the typical case systems would then perform the first action in the sequence with the highest expected return (deviations from this rule might sometimes be made to facilitate exploration). This is not just a matter of selecting actions which will tend to lead to later reward, but of selecting actions because they are favoured by specific combinations of facts about the environment, through a process which depends on representations of those facts. It is also notable that this method for selecting actions is entirely generally applicable, as a straightforward application of decision theory. MuZero learnt to play chess, Go, shogi and many Atari games using the same architecture and learning and action-selection algorithms, but it

could equally have learnt to achieve rewarding outcomes in other environments, provided that it was able to learn good models for these environments.[9]

Another way to put this argument would be to say that model-based RL systems act for reasons because they engage in instrumental reasoning which issues in action. They select actions by reasoning about which of the actions available to them will be most conducive to their goals, given the current circumstances. 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.

5. Action for reasons

Model-based RL systems select actions in a way which is interestingly different from model-free systems, and which bears notable similarities to the motivation of human actions by desires and instrumental beliefs. However, one might still question whether what model-based systems do constitutes action for reasons. Some philosophers would certainly object to this characterisation. I discuss this issue in this section, drawing on Mantel's (2017, 2018a, 2018b) theory of action for reasons to explain my own views. My reason for focusing on Mantel's theory is that it is particularly helpful for explaining my views, partly because I adopt the central element of her theory. This is the idea that to act for reasons is to exercise certain capacities, including the capacities to represent facts which count in favour of actions and to select the actions which these facts favour. The points on which I disagree with Mantel are also revealing about the substance of my views about agency in model-based RL.

Before I discuss Mantel's theory, I want to reiterate that I suspect that there are several forms of agency which are worth distinguishing for a variety of theoretical purposes. It is therefore plausible that there are forms which are more demanding than action for reasons as I understand it, for which model-based RL does not suffice, and other philosophers' accounts of what they call 'action for reasons' may identify such forms. For example, Korsgaard's (2008, 2018) view is that action for reasons requires conscious meta-cognitive reflection on one's own motives. Model-based RL does not entail this kind of reflection. But my only disagreement with Korsgaard is that I claim that there is another, less-demanding form of agency which can also reasonably be called 'action for reasons', in addition to the one she describes. The same would go for many other views on this topic.

[9] MuZero was unable to achieve good performance on some Atari games, including *Montezuma's Revenge*, which has proved particularly challenging for deep RL agents (Puigdomènech Badia et al. 2020); this is thought to be because it could not explore effectively enough, and therefore was unable to construct an adequate model of the game environment.

Mantel offers a theory of action for normative reasons in which the central idea is that this form of agency is a matter of manifesting a disposition or exercising a competence or capacity.[10] Her view is that competences and capacities are dispositions of a certain kind, so the theory can be stated in any of these terms. More specifically, her claim is that to act for a normative reason is to exercise a competence to act in the ways which are favoured by normative reasons. For example, when I add chilli oil to my soup I act for a reason because I exercise my competence to act in a way which is favoured by a fact about my present circumstances – the fact that the oil will make the soup taste better. For Mantel, it is important that I do not just have a 'brute' disposition to act in this one way in response to circumstances of this one type, but a competence to respond in the ways favoured by a variety or 'family' of different facts. This matters because it seems that an agent could have a non-rational disposition to do what is favoured by one normative reason, whereas more general or flexible responsiveness to reasons would show sensitivity to the ways in which they favour actions in context.[11]

Mantel further claims that the competence to do what is favoured by reasons is made up of three sub-competences. These are (Mantel 2018a, p. 43):

> '…the *epistemic competence* to represent the normative reasons of [some] family by descriptive beliefs, the *volitional competence* to be motivated by these descriptive beliefs to do what is favoured by the represented reasons, and the *executional competence* to execute these motivations.'

Mantel argues that each of these sub-competences is necessary, and that the unified exercise of all three is sufficient, for action for normative reasons.

A key attraction of this theory is that it seems to address apparent counterexamples to the causal theory of action for reasons in a principled, independently plausible way. Any theory of action for reasons must identify the relation between the fact which is the agent's reason for action and the action itself, in virtue of which the action is done for that reason. We can say that the fact in question is the fact that *p*. The causal theory makes the straightforward proposal that what is needed is that the agent believes that *p*, and this causes their action. The problem with this is that there are cases of deviant causal chains, in which an agent's belief causes them to act only via some interceding force which disqualifies the action from being done for the reason in

[10] Accounts of action for reasons which appeal to dispositions, competences or capacities have also been proposed by Smith (2009), Hyman (2014) and Lord (2018).
[11] See Arpaly & Schroeder 2014, p. 60 ff. for a problem case involving an isolated behavioural disposition, which they attribute to Nagel (1970) and Korsgaard (2008).

question (Davidson 1973, Arpaly & Schroeder 2015). For example, Smith (2009) describes a case in which an actor is required to tremble as though nervous, and her belief that this is required causes her to become nervous, which causes her to tremble.

One approach to addressing this problem is to add more detail to the account, in the form of extra causal steps which must be passed through as the belief causes the action. Goldman (1970, p. 62) writes that we should require that the belief causes the action 'in a certain characteristic way'. But then new counterexamples can be devised, which introduce new 'outside' forces between the new steps (Bach 1978 both proposes a revision of this kind, and notes this problem). We might also worry about the theory becoming either disjunctive or chauvinistic, because presumably different possible rational agents act through different processes. A better solution, it seems, is to appeal to dispositions, competences or capacities to act in the ways which one's beliefs favour, which are manifested or exercised in action for reasons. This works because to say that a system has a capacity to do something is roughly to say that the system has a way to do that thing. For the system to exercise the capacity is for it to do that thing in that way. So this approach allows us to cash out Goldman's idea, without specifying the form which the process of rational action selection must take, by employing concepts which are of independent philosophical interest (and have been extensively analysed, in the case of dispositions).

Turning now to the relationship between Mantel's theory and my own views, I have argued that model-based RL systems act for reasons because they represent facts that count in favour of their actions, given their goals, and exercise a general-purpose capacity to select actions which are favoured by the facts as they represent them. Capacities to represent and act on action-favouring facts are central to Mantel's theory, and the claim that the exercise of such capacities constitutes action for reasons is what allows it to enjoy the advantage just described. However, not all model-based RL systems meet all of the requirements of Mantel's theory, for reasons which are worth exploring.

The first reason is that Mantel's theory is expressed in terms of belief and motivation, and I have not attempted to justify the attribution of these forms of thought to artificial RL agents. However, my view is that a good account of action for reasons is possible without these concepts. To see this we can ask why Mantel requires the three sub-competences, rather than only the overarching competence to act in the ways favoured by normative reasons.

The value of distinguishing the epistemic and volitional sub-competences, which Mantel does by appealing to belief, has two sources. First, it ensures that the account captures the basic Dretskean insight that for an agent to act for a reason it must act because it takes or identifies the world as being a certain way. Action for reasons requires descriptive intentional states.

Second, it helps the account to solve what Mantel (2018b) calls 'Davidson's challenge', which is the problem of distinguishing the reasons for which an agent acts, when there are multiple reasons which count in favour of their action. The appeal to belief makes it possible, to a greater extent, to trace the path from particular reasons to actions. An appeal specifically to belief is not necessary for these functions, however, because a less specific requirement for descriptive representation would serve equally well. To use descriptive representations in instrumental reasoning which issues in action is to take the world to be a certain way, whether or not the representations amount to beliefs, and such representations are equally suitable for meeting Davidson's challenge.

Mantel's distinction between volitional and executive sub-competences and appeal to motivation, in contrast, seems to be explained by features of human agency for which some artificial RL systems have no analogues. Mantel (2018a, p. 17) writes that she uses the term 'motivation' in a broad sense, interchangeably with 'desire', and that intention is for her a form of motivation. It makes sense to employ these concepts in the human context because human actions are relatively loosely connected to some of the model-based practical reasoning which we engage in. Humans reason about how to act in advance, for situations which we expect to encounter; our agency is hierarchical, in the sense that we sometimes perform whole sequences of 'lower-level' actions guided by a single choice made at a higher level; and executing the actions we have selected is sometimes challenging. So volition and execution can be seen as distinct stages in the process. These three phenomena are all possible for artificial RL systems but absent in the most basic cases of model-based RL. Action for reasons is still possible in cases in which they are absent, however – someone might act for a reason in performing a very simple action, like pressing a button, in response to a situation which they did not anticipate. So I see no need for this distinction in the account. We can just as well think of action for reasons as involving two capacities – to represent action-favouring facts and to select the actions they favour – rather than three.

Despite these points, many philosophers may still object that artificial RL systems lack mental states and that this disqualifies them from acting for reasons. In particular, such systems arguably lack desires or volitions, so while they may select outputs which are favoured in some sense by facts about their environments, they are very different from 'real agents'. There is nothing that they want or care about.

I agree that there are some model-based RL systems which lack desires. In the context of the claim that the human mind implements RL algorithms, empirically-informed accounts of desire have suggested that desires are representations of state values which we use when engaging in

model-based RL (Cushman & Paul 2022, Railton 2012, Butlin 2017). But this implies only that a role in model-based RL is necessary for desire, not that it is sufficient. Desires also have a phenomenological role, especially in connection with affective phenomenology, and human and animal desires are connected to our biological drives. Chang (2004) and Shaw (2021) appeal to these features in accounts of the significance of desire for action for reasons, and they are likely to be absent in many or all artificial RL systems.

It may well be that the distinction between agents which have desires and agents which lack them – or some similar distinction – is a very important one. Once again, my assumption is that there are a variety of forms of agency which are significant in different ways. Some model-based RL systems are controlled by very simple programs and operate in very simple environments, so it is to be expected that model-based RL is not sufficient for some of the most important forms. However, the point that some artificial RL systems lack desires does not seem to me to undermine the arguments that I have made so far in this paper. RL systems in general pursue goals through their interaction with their environments, learning to produce outputs selectively for their instrumental value. This is sufficient grounds to attribute minimal agency. And model-based systems go beyond this by learning facts which count in favour of actions, given their goals, and employing general-purpose methods to select the actions which these facts favour, so since they are agents, there is a substantive sense in which they act for reasons.

A second respect in which model-based RL fails to meet the requirements of Mantel's theory is that she offers it as an account of action for *normative* reasons, and has a demanding conception of reasons of this kind. She writes that a normative reason is 'a fact that objectively favors an action given the right ethical theory' (2018b, p. 208) and gives examples showing that on her conception facts can favour agents' actions, given their goals, while not being normative reasons for action. One example involves a killer who shoots their victim a second time because they are still alive. The fact that the victim is still alive may be the 'agent's reason' for shooting, she writes, but would not be a normative reason because it does not 'objectively favour' the action. She also develops her theory partly in an attempt to understand morally worthy actions, for which agents deserve credit (2018a). This aspect of Mantel's views is significant because an artificial RL system might operate in a virtual environment in which none of its actions matter (including to itself). By Mantel's standards, such a system might have no normative reasons for action, so it would not be capable of acting for such reasons.

I differ from Mantel here because I am less interested in understanding agents' sensitivity to facts which objectively favour actions, independently of their goals, and more interested in action for reasons more broadly construed. A broader construal is important because it allows us to

recognise the features in common between the action of Mantel's killer, that of a person who acts for a reason in pursuing a morally neutral goal such as winning a judo bout, and that of an animal which acts for a reason when it forages in a particular location. One way to understand reasons in this sense is to draw on Finlay's (2014) analysis of normative language, including the word 'reason'. Simplifying somewhat, Finlay argues that the expression 'a reason for *s* to ϕ' means *an explanation why it is good, relative to some end, for s to ϕ* (ch. 4). In this sense there can be reasons relative to any end, worthy or otherwise. To put it another way, my interest is in reasons understood as facts that count in favour of actions, relative to goals. If all model-based RL systems have goals, it follows that they have reasons for action in this sense.

In short, I agree with Mantel that action for reasons should be analysed in terms of exercises of capacities or competences, that agents must represent their reasons, and that the capacities in question must be sufficiently general. However, her view can be modified, while retaining its main advantages, by removing the references to belief and motivation and broadening the conception of normative reasons. This yields an analysis of action for reasons which is more inclusive but remains substantive.

6. Conclusion

Both model-free and model-based RL systems learn to produce outputs for their instrumental value. That is, they learn to exploit the effects that their outputs have on their environments, and thus on subsequent inputs, to achieve good performance over episodes of interaction. By learning to do this these systems come to pursue goals and to meet standards for minimal agency. Both model-free and model-based RL systems represent their immediate circumstances and store information which allows them to select reward-conducive actions, but model-based systems have the special feature that they model the transition function and use the representations making up this model in instrumental reasoning. This means that they represent and act on the basis of facts which count in favour of their actions, given their goals, and therefore that they act for reasons.

References

Arpaly, N. & T. Schroeder. 2014. *In Praise of Desire*.

Arpaly, N. & T. Schroeder. 2015. A causal theory of acting for reasons. *American Philosophical Quarterly* 52 (2): 103-114.

Artiga, M & M. Martínez. 2016. The organizational account of function is an etiological account of function. *Acta Biotheoretica* 64 (2): 105-117.

Bach, K. 1978. A representational theory of action. *Philosophical Studies* 34: 361-379.

Barandiaran, X., E. Di Paolo & M. Rohde. 2009. Defining agency: Individuality, asymmetry, normativity and spatio-temporality in action. *Adaptive Behavior* 17: 367-386.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*.

Bratman, M. 2000. Valuing and the will. *Philosophical Perspectives* 14: 249-265.

Burge, T. 2009. Primitive agency and natural norms. *Philosophy & Phenomenological Research* 79: 251-278.

Butlin, P. 2017. Why hunger is not a desire. *Review of Philosophy & Psychology* 8: 617-635.

Butlin, P. 2022. Machine learning, functions and goals. *Croatian Journal of Philosophy* 22: 351-370.

Chang, R. 2004. Can desires provide reasons for action? In Wallace et al., eds., *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*.

Cushman, F. & L. A. Paul. 2022. Are desires interdependent? In Doris & Vargas, eds., *Oxford Handbook of Moral Psychology*.

Davidson, D. 1973. Freedom to act. In Honderich, ed., *Essays on Freedom of Action*.

Dennett, D. 1996. *Kinds of Minds*.

Dolan, R. & P. Dayan. 2013. Goals and habits in the brain. *Neuron* 80 (2): 312-325.

Dretske, F. 1985. Machines and the mental. *Proceedings and Addresses of the American Philosophical Association* 59: 23-33.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*.

Dretske, F. 1993. Can intelligence be artificial? *Philosophical Studies* 71(2): 201-216.

Dretske, F. 1999. Machines, plants and animals: The origins of agency. *Erkenntnis* 51: 523-535.

Finlay, S. 2014. *Confusion of Tongues: A Theory of Normative Language*.

Garson, J. 2019. *What Biological Functions Are and Why They Matter*.

Gershman, S. & N. Daw. 2017. Reinforcement learning and episodic memory in humans: An integrative framework. *Annual Review of Psychology* 68: 101-128.

Glock, H.-J. 2009. Can animals act for reasons? *Inquiry* 52: 232-254.

Goldman, A. 1970. *A Theory of Human Action*.

Haas, J. 2022. Reinforcement learning: A brief guide for philosophers of mind. *Philosophy Compass* 17 (9).

Halina, M. 2021. Insightful artificial intelligence. *Mind & Language* 36: 315-329.

Hofmann, F. & P. Schulte. 2014. The structuring causes of behavior: Has Dretske saved mental causation? *Acta Analytica* 29: 267-284.

Hyman, J. 2014. Desires, dispositions, and deviant causal chains. *Philosophy* 89 (1): 83-122.

Jamieson, D. 2018. Animal agency. *The Harvard Review of Philosophy* 25: 111-126.

Kagan, S. 2019. *How to Count Animals, More or Less*.

Kenton, Z. et al. 2022. Discovering agents. *arXiv* preprint.

Kool, W., F. Cushman & S. Gershman. 2018. Cooperation and competition between multiple reinforcement learning systems. In Morris, Bornstein & Shenhav, eds., *Understanding Goal-Directed Decision-Making: Computations and Circuits*.

Korsgaard, C. 2008. *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*.

Korsgaard, C. 2018. *Fellow Creatures*.

Krizhevsky, A., I. Sutskever & G. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* 60: 84-90.

Lord, E. 2018. *The Importance of Being Rational*.

Mantel, S. 2017. Three cheers for dispositions: A dispositional account of acting for a normative reason. *Erkenntnis* 82: 561-582.

Mantel, S. 2018a. *Determined by Reasons: A Competence Account of Acting for a Normative Reason*.

Mantel, S. 2018b. Because there is a reason to do it: How normative reasons explain action. *Analytic Philosophy* 59 (2): 208-233.

Meincke, A.-S. 2018. Bio-agency and the possibility of artificial agents. In D'Aragona et al., eds., *Philosophy of Science: Between the Natural Sciences, the Social Sciences, and the Humanities*.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*.

Mnih, V. et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (7540): 529-533.

Mossio, M., C. Saborido & A. Moreno. 2009. An organizational account of biological functions. *The British Journal of Philosophy of Science* 60 (4): 813-841.

Nagel, T. 1970. *The Possibility of Altruism*.

Niv, Y. 2009. Reinforcement learning in the brain. *Journal of Mathematical Psychology* 53: 139-154.

Omohundro, S. 2008. The basic AI drives. In Wang, Goertzel & Franklin, eds., *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*.

Papineau, D. 2001. The evolution of means-end reasoning. *Royal Institute of Philosophy Supplement* 49: 145-178.

Puigdomènech Badia, A. et al. 2020. Agent57: Outperforming the Atari human benchmark. *arXiv* preprint.

Railton, P. 2012. That obscure object, desire. *Proceedings and Addresses of the American Philosophical Association* 86: 22-46.

Schlosser, M. 2012. Taking something as a reason for action. *Philosophical Papers* 41(2): 267-304.

Schrittwieser, J. et al. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* 588: 604-609.

Schultz, W., P. Dayan. & P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* 275: 1593-1599.

Shaw, A. 2021. Desire and what it's rational to do. *Australasian Journal of Philosophy* 99 (4): 761-775.

Shea, N. 2018. *Representation in Cognitive Science.*

Shulman, C. & N. Bostrom. 2021. Sharing the world with digital minds. In Clarke, Zohny & Savulescu, eds., *Rethinking Moral Status.*

Sebo, J. 2017. Agency and moral status. *Journal of Moral Philosophy* 14: 1-22.

Silver, D. et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529 (7587): 484-489.

Silver, D. et al. 2018. A general reinforcement learning algorithm that masters chess, shogi and Go through self-play. *Science* 362: 1140-1144.

Smith, M. 2009. The explanatory role of being rational. In Sobel & Wall, eds., *Reasons for Actions.*

Sutton, R. 1991. Dyna, an integrated architecture for learning, planning and reacting. *SIGART Bulletin* 2 (4): 160-163.

Sutton, R. & A. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd edition).

Velleman, D. 2000. *The Possibility of Practical Reason.*

Wallach, W. & C. Allen. 2008. *Moral Machines: Teaching Robots Right from Wrong.*

Whittlestone, J., K. Arulkumaran & M. Crosby. 2021. The societal implications of deep reinforcement learning. *Journal of Artificial Intelligence Research.*