Croatian Journal of Philosophy
Vol. XXII, No. 66, 2022
https://doi.org/10.52685/cjp.22.66.5
Received: August 10, 2022
Accepted: September 22, 2022

Machine Learning, Functions and Goals

PATRICK BUTLIN
University of Oxford, Oxford, UK

Machine learning researchers distinguish between reinforcement learning and supervised learning and refer to reinforcement learning systems as "agents". This paper vindicates the claim that systems trained by reinforcement learning are agents while those trained by supervised learning are not. Systems of both kinds satisfy Dretske's criteria for agency, because they both learn to produce outputs selectively in response to inputs. However, reinforcement learning is sensitive to the instrumental value of outputs, giving rise to systems which exploit the effects of outputs on subsequent inputs to achieve good performance over episodes of interaction with their environments. Supervised learning systems, in contrast, merely learn to produce better outputs in response to individual inputs.

Keywords: Agency; machine learning; reinforcement learning; artificial intelligence; Dretske.

1. *Introduction*

One of the most powerful ideas in modern philosophy of mind is that an entity's origins can ground standards of success or evaluation to which its activities are subject. The relevant origins here are histories of learning or selection. This idea builds on the claim from philosophy of biology that selective history grounds biological function (Garson 2016) and has been prominently used in theories of representation (e.g. Millikan 1984, Papineau 1993, Shea 2018), as well as in teleofunctional theories of mental states (Sober 1985, Lycan 1987). In the theory of representation this idea helps to explain the correctness conditions which are deeply connected with meaning. In teleofunctionalism it helps to explain the fact that mental states and processes stand in normative

relations to one another—for instance, that it is part of the function of desires to cause motivation to act in combination with beliefs.

This idea may also be used in analysing agency. Agents engage in activity which is purposeful, functional, or otherwise governed by norms or standards, and their etiologies may ground these features. Glaciers interact with their environments but they are not agents because this activity is not governed by standards of correctness or evaluation. There is no sense in which glaciers aim to, or are supposed to, meet any such standards. Living organisms, in contrast, are at least candidates for agency, because much of their activity is purposeful or functional. I will say that agents engage in "norm-governed" activity, using the word "norm" very broadly to refer to non-arbitrary standards of correctness or of better or worse performance. Norm-governed activity is a necessary but not sufficient condition for agency, because the heart—for example—engages in activity which can be more or less successful according to its biological function, but the heart is not an agent. So agency is a species of which norm-governed activity is the genus.

Another way to see the point that agency is norm-governed is to start from the idea that agents pursue goals. If this is the case, agents' activity can be evaluated according to whether it helps to achieve their goals. Having a goal and having a function are two different ways to be subject to norms. In this paper, I will suggest that to have a goal, and thus to be an agent, it is necessary to have a history of learning or selection of a particular kind. Histories of this kind are made possible by certain capacities, and make others possible in turn. I will focus on formulating my claim in the context of a particular case; more work will remain to test the claim in other contexts.

My discussion will focus on the case of machine learning and in particular on the distinction between reinforcement learning and supervised learning. Machine learning researchers standardly refer to entities which undergo reinforcement learning as "agents", and reinforcement learning algorithms are designed to solve problems of the same general form of those which face biological agents (Sutton and Barto 2018). Furthermore, concepts and algorithms from reinforcement learning research are now widely used to explain value learning and action selection in humans and other animals (Niv 2009, Dolan and Dayan 2013). So it is natural and plausible to associate reinforcement learning with agency. In contrast, there are many systems trained by supervised learning, such as image classifiers, spam filters and translation tools, which do not seem to be agents. I will suggest an account of agency which vindicates these initial impressions, on the grounds that reinforcement learning is an example of the kind of process which gives rise to goals, but supervised learning is not.

This paper is therefore concerned with minimal agency—with the most basic distinction between those entities which are agents and

those which are not. It contrasts with much philosophical research on agency, which is concerned with the subtleties of human agency. Humans make plans, collaborate with others, experience emotions, and reflect on our own motives and choices, but none of these features seems to be essential to agency. I will start from a theory of minimal agency developed by Fred Dretske (1985, 1988, 1993, 1999), partly because it is abstract enough to be applied to the cases I am concerned with. I will set aside alternative approaches to minimal agency which are more specifically focused on the biological domain, such as those by Barandiaran et al. (2009) and Burge (2009).

In much of the paper I will not discuss the normative aspect of agency explicitly. After presenting Dretske's theory I will criticise it on the grounds that it implies that supervised learning-trained image classifiers are agents (section 2). I will then examine the differences between supervised learning and reinforcement learning, and propose a modification to Dretske's account, in section 3. In section 4 I will illustrate and elaborate on my proposal by discussing further examples of machine learning. In section 5, however, I will return to the idea that agency arises from histories of a particular kind, which give rise to entities which have goals and are consequently subject to associated norms. I will reformulate my proposal in these terms, building on the claim that natural selection gives rise to traits with biological functions.

2. *Dretske's theory of agency*

According to Dretske (1993, 1999), action is behaviour "controlled" or "governed" by thought. His account of agency forms part of his ambitious and elegant theory of intentionality and mental causation, which is presented in *Explaining Behavior* (1988) and several associated articles. One central claim of the account is that learning is necessary for agency. This learning must establish a structure in which a form of behaviour is produced selectively in response to features of the environment, through the operation of an internal state of the system. This internal state must be correlated with a feature of the environment, and must cause the behaviour partly in virtue of this correlation. That is, for some output of a system of type B to be an action, the following conditions must be met:

i. Internal states of the system of some type R are correlated with a feature of the environment E.
ii. The system learns to produce outputs of type B when in R-states.
iii. This learning happens in part because R-states are correlated with E.

For a system as a whole to be an agent, it must perform actions; a token output of type B is an action when it is caused by an internal state of type R through the route established by learning.

For example, consider a bird which learns to eat red pellets. For this to happen, the bird must have a visual system which enters a state of a certain kind when red pellets are in its field of view. If it pecks at and eats red pellets in the course of exploring its environment, and this behaviour is rewarded (e.g. because the pellets are palatable), it may learn to eat them selectively. This will involve a causal connection being formed between the visual system state that is correlated with red pellets and the behaviour of pecking and eating. This process will result in an arrangement which satisfies Dretske's conditions, and hence, according to Dretske, in the bird's becoming disposed to perform the action of eating red pellets.

In this case, the visual system state would not merely carry information about the presence of red pellets, but would come to be used as an indicator of red pellets. For Dretske, this means that it would come to represent the presence of red pellets. Alternatively, as he also puts it, it means that being in this internal state would amount to the bird's "thinking", or "believing", that red pellets are before it.

This "thought" or "belief" would then cause the behaviour of pecking and eating. For Dretske, a crucial point is that it would cause this behaviour in virtue of its content (behaviour being caused by thought is not enough, because this could happen without content being relevant). This would be the case because the correlation between the state and the presence of red pellets—the relation that underlies content—would have been a contributing cause of the connection's being established between the state and the behaviour. We would have a case of behaviour governed by thought, and therefore of agency.

As an influential account of content, mental causation and agency this picture has naturally been criticised.[1] One important criticism offered by Dennett (1991) is that it is not clear why the relationship between environmental conditions, internal states and behaviours must be established by learning rather than by evolution or design. A simple but unsatisfying response is that plants and simple artifacts would count as agents without the learning requirement. Thermostats are constructed so as to have internal states which correlate with low temperatures, which cause heating-activation outputs. The scarlet gilia, a plant which Dretske (1999) uses as an example, has flowers which change colour at the height of summer. It must therefore have some internal state which is correlated with the season, which is a proximal cause of this change. But in neither case is it appealing to say that the system is an agent, or that its output is "governed by thought". Some further justification might be achieved by saying that agents must be "autonomous" in the sense of Russell and Norvig (2010)—that is, that they must have a degree of independence from the knowledge of their designers, or more generally from the information which contributed to their initial forms. There is more to be said to fully justify the learning

[1] For criticisms which I will not discuss here, see Hofmann and Schulte (2014).

requirement, but here I will grant Dretske the point, in order to concentrate on a different feature of his theory.

I claim that Dretske's theory is insufficiently demanding because it entails that certain supervised learning-trained systems are agents.[2,3] For example, consider AlexNet (Krizhevsky et al. 2012), an image classifier using a deep convolutional neural network which was one of the defining advances of the development of deep learning. AlexNet is trained to label images as belonging to one of 1000 categories, in the following way. First an image is drawn from the training set and given to AlexNet as an input. This causes the network to produce some output, which takes the form of an assignment of probabilities to each of the categories. The correct label is provided, and the system uses gradient descent and backpropagation to adjust the network weights. This process is then repeated with further images from the training set, and the adjustments gradually increase the likelihood that the network will assign the highest probability to the correct label.

This may reasonably be described as a process of learning. The system undergoes endogenous, systematic changes in response to feedback which improve its performance, and it does so because it has been designed to change in this way. Furthermore, this learning seems to result in a situation which satisfies Dretske's criteria. After it has received some training, patterns of node activation in AlexNet will be correlated with type of input image—there may be some particular pattern correlated with images of pandas, for example. These patterns will cause AlexNet to produce particular kinds of outputs. The "panda" pattern will cause outputs which assign high probability to the "panda" category, and low probability to other categories. This situation will arise because the "panda" patterns are correlated with images of pandas, so weight combinations through which these patterns cause "panda" outputs will tend to be preserved. So AlexNet learns to produce outputs selectively in response to features of its environment, via internal states which indicate these features.

This is a problem for Dretske's account because AlexNet does not pursue any goal, and is not naturally described as an agent. It performs the function of classifying images, but not every entity which performs a function is an agent (as illustrated by the case of the heart). In the next section I will contrast supervised learning with reinforcement learning, which will allow me to give a more detailed analysis of this case.

[2] Strikingly, Dretske (1993) writes that genuine artificial intelligence is impossible, because being artificial is incompatible with being a product of learning, and the latter is necessary for genuine intelligence. This is surprising because he mentions learning in connectionist systems in *Explaining Behavior*.

[3] "Systems" here refers to particular implementations of algorithms—in this case, algorithms generated by the operation of implementations of further, supervised learning algorithms. Throughout this paper, when I suggest that artificial systems could be agents, my claim concerns implementations, not algorithms.

3. *Supervised learning, reinforcement learning and agency*

Machine learning problems and techniques are generally taken to belong to one of three classes: unsupervised learning, supervised learning and reinforcement learning. I consider only the latter two here, leaving unsupervised learning aside. In this section I describe supervised learning and reinforcement learning, then identify a difference which matters for agency.

According to Russell and Norvig's standard textbook on AI (2010: 695),

> The task of supervised learning is this:
>
> Given a training set of N example input-output pairs
>
> $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N),$
>
> Where each y_j was generated by an unknown function $y = f(x)$, discover a function h which approximates the true function f.

AlexNet is an example of a solution to a task of this form, because there is some function which takes each image in a labeled set to the correct label. An artificial neural network such as AlexNet can be seen, at each stage of training, as realising whatever function describes the transitions it is disposed to make from inputs to outputs. As AlexNet is trained this function comes to more closely approximate the true, target function.[4]

There are two noteworthy features of supervised learning which help to distinguish it from reinforcement learning. These both arise from the form of the training set, as a non-ordered set of input-output pairs. First, the feedback which the learning system receives, which drives its learning, specifies the correct output for the input just provided. Second, the input which is provided on each occasion and the correct output for that input are independent of any other actual or potential inputs or outputs. In particular, the system's outputs do not affect subsequent inputs.

Russell and Norvig define reinforcement learning as follows (2010: 830):

> The task of reinforcement learning is to use observed rewards to learn an optimal (or near optimal) policy for the environment.

Rewards are a form of feedback in which a numerical signal, which can have a positive, negative or zero value, is provided to the learning system after it produces each output. In reinforcement learning the next input (which is called a "state") depends probabilistically on the previous one and the system's output (called an "action"). The optimal policy for the environment is defined as that which maximises expected cumulative reward.

[4] For more on convolutional neural networks such as AlexNet, see Buckner (2019).

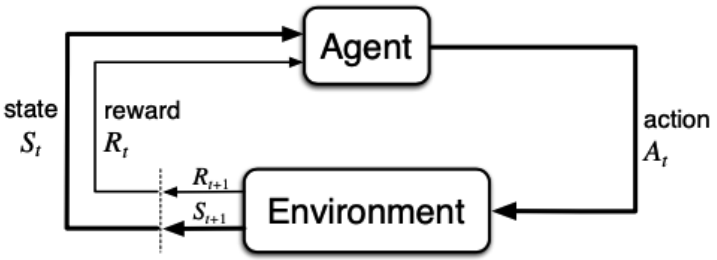

Figure 1. Illustration of reinforcement learning from Sutton and Barto (2018).

This arrangement is illustrated in figure 1. Here the "agent" is the system which undergoes reinforcement learning. At each time-step the system receives the state of the environment as input, produces an action as output, and receives a reward and an observation of the new state. In reinforcement learning environments are made up of transition functions, which describe the probabilities of new states given prior states and actions, and reward functions, which describe how much reward the agent will receive after each action.

An important advance in reinforcement learning from roughly the same period as AlexNet combined deep neural networks with a method called Q-learning to achieve human-level performance on Atari games (Mnih et al. 2015).[5] We can call this system DQN (for "Deep Q-Network"). As all reinforcement learning systems do, DQN receives both observations of the state of the environment and reward. Observations of the state of the environment take the form of maps of pixel values making up what would be displayed on a screen for human players, and reward is constituted by the game score. Outputs are actions possible for human players, such as producing the in-game effect of pressing a joystick button. DQN is trained separately on each game, losing its capacity to play one when trained on another.

To understand how DQN works, the most important element is the Q-learning algorithm. The function *Q(s, a)* is the action value function for the environment, describing how much cumulative reward can be expected to follow from taking action *a* in state *s* (which also depends on the system's policy, i.e. the actions it will subsequently choose). This function is somewhat analogous to the target function *f* in supervised learning, in that a reinforcement learning system will behave optimally if it always selects the action that maximises the Q-function for the current state. Analogously to AlexNet, DQN's outputs are determined by maximising its current estimate of the Q-function. There is the significant difference, though, that DQN is not given the true Q-value for the action it has just taken. Instead, it observes only the immediate change in the game score. This is very different, because—for example—an ac-

[5] The description given here is simplified in significant respects; see Mnih et al.'s paper for more details.

tion may cause no immediate change in the score, and yet be necessary to reach a state from which the highest scores are accessible.

Nonetheless, it is possible to use reward feedback to reach an approximation of the true Q-function. The method is to update estimated Q-values in the direction of the temporal difference error, given by the following formula:

$$R + \gamma\, Q(s', a') - Q(s, a)$$

Here R is the reward, γ is a discount factor, $Q(s', a')$ is the estimated value of the best action in the new state, and $Q(s, a)$—the value to be updated—is the estimated value of the action just taken in the previous state. The effect of this is that credit for rewards is passed back through the sequences of actions that lead to them.

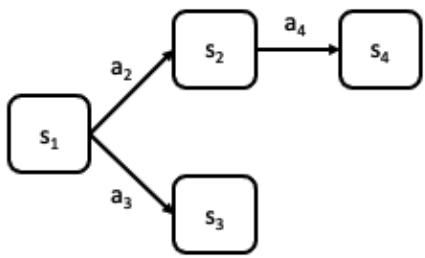

Figure 2. Illustration of Q-learning.

For example, consider the partial environment shown in figure 2, and suppose that the agent receives a high reward in s_4. In that case the temporal difference error for (s_2, a_4) is likely to be positive, so the agent's estimate of $Q(s_2, a_4)$ will be adjusted upwards. When the agent next performs a_2 in s_1, and thus reaches s_2, this higher value of $Q(s_2, a_4)$ will again likely mean a positive temporal difference error—because this will take the place of $Q(s', a')$ in the formula—so the agent's estimate of $Q(s_1, a_2)$ will be boosted. Credit for getting the high reward will be distributed back from a_4 to a_2 (and it could continue to be passed back in the same way). This could lead to the system forming a disposition to perform a_2 rather than a_3 in s_1 even if the latter led to greater immediate reward. In this way, the actions in sequences which lead to high rewards come to be represented as having high Q-values.

Reinforcement learning differs from supervised learning in each of the two features mentioned above. First, the feedback which drives reinforcement learning does not specify the correct output for the input just received. Instead, it is made up of an observation of the next state and a reward signal. Second, the identity of the next state is not independent of the previous one—instead, it is affected by the previous state and the action just performed. This means that reinforcement

learning systems engage in interaction with their environments—these are not just sources of inputs to which they must respond, but are also affected by their outputs in ways which affect their inputs in turn. In addition to these two features, in reinforcement learning there is a measure of success over episodes of interaction, and systems are equipped with algorithms which promote good performance on this measure. To perform well, in general, a reinforcement learning system must do more than just learn which actions yield most immediate reward. It must also learn how to reach states from which high levels of reward are available. That is, it must learn to exploit the fact that its outputs affect which inputs it will receive.

I propose that reinforcement learning systems are agents because, in addition to satisfying Dretske's conditions, they are capable of *instrumental* behaviour. To behave instrumentally is to produce outputs because these outputs contribute to good performance over episodes of interaction, such as by making it possible to access later rewards. Instrumental behaviour is both possible and necessary for reinforcement learning systems for the reasons just described. In particular, Q-learning and related methods produce instrumental behaviour because outputs come to be selected in virtue of their conduciveness to later rewards.

In contrast, AlexNet's outputs cannot be instrumental because they have no effect on subsequent inputs. Even if they did have an effect, the learning method employed in AlexNet is not sensitive to sequences of inputs, outputs and subsequent inputs, so it could not learn to engage in instrumental behaviour. The gradient descent algorithm by which AlexNet's weights are adjusted works by comparing the actual output for the current input with the correct output for that input. The feedback in supervised learning—that is, the information provided to the system which is affected by its outputs and which drives learning—does not include the identity of the next input. This difference between AlexNet and reinforcement learning systems makes sense because for AlexNet good performance overall just consists in producing the correct output for each input. For reinforcement learning systems, what makes outputs correct is how they contribute to maximising reward.

This view can be captured by the following claim about agency:

Instrumental view: An entity is an agent if and only if:

i. It produces some of its outputs selectively in response to inputs, as a result of a process which includes learning.

ii. This process is sensitive to instrumental value, where this means that it is influenced by information about input-output-input contingencies and functions to promote a specific form of feedback over episodes of interaction with the environment.

This view of agency combines two features: instrumentality in behaviour, and the learnt selectivity which Dretske describes. These two

features appear to be orthogonal, in that AlexNet learns to produce outputs selectively, but these are not instrumental, whereas a robot programmed to move efficiently through a specific maze would produce instrumental outputs without learning. However, it would be a mistake to think of my account as made up of separate instrumentality and learnt-selectivity conditions. Instead, what is crucial for agency is that the learning process is sensitive to instrumental value, so the system learns to produce outputs selectively because they contribute to good performance over an episode of interaction. One of the examples I will discuss in the next section will serve to illustrate this point.

4. *More on machine learning*

In this section I will discuss a series of further examples involving machine learning. Subsections 4.1 and 4.2 will cover other varieties of reinforcement learning, and add more detail to my account of how this form of learning is related to agency. Subsection 4.3 will discuss the possibility of using supervised learning to mimic optimal behaviour in a reinforcement learning-style environment; this case will prompt the clarification to my view suggested at the end of the last section. Finally, in subsection 4.4 I will comment briefly on agency in large language models.

4.1 *Varieties of reinforcement learning*

In the theory of reinforcement learning, a distinction is often made between "model-free" and "model-based" methods. The difference is that model-based methods involve the system learning and using a representation of the transition function, which can also be thought of as a model of the environment. Q-learning is a typical example of temporal difference learning, which is the most broadly-applicable form of model-free reinforcement learning. So in this subsection I will comment on varieties of reinforcement learning other than temporal difference learning, beginning by showing that systems which use typical model-based methods satisfy Dretske's conditions for agency and are capable of instrumental behaviour.

This claim can be illustrated by considering a model-based algorithm called R-Max (Brafman and Tennenholtz 2002). In this algorithm, look-up tables are maintained which store estimates of the transition function and reward function for the environment (the use of look-up tables means that this method is only suitable for finite environments). The rows in the transition function table record information about the new state which is expected following each action in each initial state, and the rows in the reward function table record the reward which is expected in each state. Actions are selected by exhaustive calculation of the cumulative reward of their expected consequences, looking ahead a fixed number of steps, with the action that begins the most rewarding sequence being chosen.

A system using this algorithm would satisfy Dretske's conditions because it would produce outputs selectively as a result of learning. After an initial period of exploration, such a system would develop dispositions to perform particular actions in particular states because its model would imply that these would lead to the greatest cumulative reward over the period to which its look-ahead extended. These actions would be caused by internal states correlated with states of the environment, and the causal links between internal states and actions would be explained by a combination of learning—which would establish the agent's model of the transition function and reward function—and reasoning—which would be used to select actions on the basis of the model.

Furthermore, the system would be capable of instrumental behaviour, because it would look ahead more than one step when selecting outputs. It would choose the actions which would allow it to maximise cumulative reward over multiple steps, meaning that its actions would be chosen for their contributions to good performance over episodes of interaction. The cases of temporal difference learning and model-based reinforcement learning illustrate that instrumental behaviour can be generated in different ways—either through learning algorithms which carry information about reward backwards through sequences of actions, or through action selection algorithms which use learnt models to look forward through such sequences.

A different form of model-free reinforcement learning is called Monte Carlo control (Sutton & Barto 2018). Monte Carlo control is notable because, whereas the sensitivity to instrumental relationships between actions and subsequent states is more explicit in R-Max than in Q-learning, this sensitivity is even less explicit in Monte Carlo control than in Q-learning. Monte Carlo control works in the following way. The system's purpose is to maximise reward in an environment with an end-state, which it engages with repeatedly (Monte Carlo control only works in cases like this). It starts by following some fixed policy many times, perhaps from a range of initial states. It records how much total reward it receives subsequent to each state-action pair on each occasion, then estimates Q-values for the policy it has been following by taking the mean of each set of observations. Then it improves its policy by choosing actions with higher Q-values, and repeats the process.

Monte Carlo control involves learning to select outputs for their contributions to cumulative reward, and hence involves exploiting the fact that outputs affect subsequent inputs. However, it does not depend on the agent's representing which states its actions lead to—either to feed into immediate updates as in Q-learning, or as part of the process of constructing a model. Instead, which states actions lead to influences how the system is updated by affecting the cumulative reward that follows actions. In this way, systems using this method are influenced by

information about instrumental relationships, so Monte Carlo control is sufficient for agency.

However, systems designed to solve two other problems studied in the context of reinforcement learning are not generally agents. These are the problem of planning, and multi-armed bandit problems (Sutton and Barto 2018). Planning is using a model of an environment which has been provided by the programmer to find an optimal policy. Planning is a crucial element of model-based reinforcement learning, but the capacity to plan does not suffice for agency, because it does not involve learning. Planners have little autonomy.

Multi-armed bandit problems are problems in which a number of outputs ("actions") are available to a system, each of which leads stochastically to a range of rewards, so that the system must learn which action is most rewarding. Systems for solving bandit problems are not generally agents, however, because the state of the environment does not change. So learning quickly about the relative values of outputs and maximising cumulative reward does not depend on exploiting the effects of outputs on subsequent inputs.

4.2 *Reinforcement learning systems pre- and post-training*

A further feature of reinforcement learning systems which calls for clarification of my account is that they change over time. Their abilities to navigate particular environments are gained only gradually, with this process often starting from an initial condition in which they select outputs randomly. In addition to this, engineers sometimes train systems with reinforcement learning only up to the point at which they reach a certain level of performance. After this the systems operate in the environment using a fixed policy or model, learnt during the training phase.

Different approaches to theorising about agency would give different verdicts on the status of reinforcement learning systems pre- and post-training. An approach which distinguished agents from non-agents according to whether they have the capacity to learn to behave in the relevant way would claim that pre-training systems are already agents, but systems which have been "frozen" after training are agents no longer. Combining this approach with my proposal that sensitivity to instrumental value matters would yield the view that agents are those entities with the capacity to learn to produce outputs selectively for their instrumental value. However, an alternative approach might claim that agents are those entities which perform actions, and actions are those outputs which are caused in the right way. Although the former approach has some attraction, I favour the latter. For an output to be an action it must be produced because the system has undergone a process which includes learning and is sensitive to instrumental value. This entails that reinforcement learning systems become agents gradually as they learn, because learning gradually comes to play a greater

role in explaining their outputs. It also entails that post-training systems which can no longer learn are still agents, because they still produce outputs as the result of a process of the right kind.

This approach has two advantages. First, as I will explain further in section 5, it makes it possible to analyse agency as a form of norm-governed activity, with the existence of the relevant norms grounded in history. Second, it is based on an analysis of actions as outputs which are caused in a certain way, and therefore subject to a certain form of explanation. It makes sense to use an account of action as the basis for a theory of agency, both because their capacity to perform actions is what is interesting about agents, and because not all outputs of agents are actions, so a substantive theory of action is needed in any case.

It may be objected at this point that I have not considered the possibility of an account of action which is based on proximal causes, such as reasoning which takes place "in the moment", rather than on the more distal role of learning. An account of this kind would avoid the potentially troubling implication of my view that a relatively long history is required.[6] One problem with accounts of this kind, however, is that they seem to have trouble distinguishing between AlexNet and DQN. Neither does much reasoning about which output to produce in response to a given input, but they still produce their outputs for very different reasons, and closer inspection of these shows important commonalities between DQN and model-based systems which do engage in in-the-moment reasoning.

4.3 *Mimicing agents using supervised learning*

It is sometimes argued that reinforcement learning is not necessary for agency on the grounds that it is possible to train a system by supervised learning that will mimic the behaviour of any reinforcement learning agent. The optimal policy for an environment is a function from states to actions, so if we know this function we can train a system to approximate it by supervised learning. More generally, if we know how a given reinforcement learning system will behave in a given environment, we can describe its behaviour as a function from states to actions, and again use supervised learning to train a system to mimic it. I claim that supervised learning systems of this kind are not agents, because—as I have just argued—the status of an entity as an agent depends on its history, not just on its current dispositions.

[6] A theory according to which a history of learning is required for agency faces the objection that a "swampman"—that is, a perfect replica of a living, adult human which emerges by chance from a swamp—would not immediately be an agent. I think this is the correct verdict on this case (see e.g. Millikan 1996, Shea 2018). See also McKenna (2016) and Zimmerman (2003) for discussions of other aspects of the role of history in agency.

This case is notable because it shows that learnt selectivity and instrumentality need to be combined in the right way to yield an attractive theory of agency. Dretske's theory entails that the status of an entity as an agent depends on its history because it requires an agent's dispositions to be a product of learning. However, we have already seen that Dretske's theory entails that supervised learning systems can be agents, so appealing to this theory alone will not justify a denial of agency in the present case. In addition to this, there is a sense in which the supervised learning "mimic" performs outputs for their instrumental value, because it is this value that explains why the reinforcement learning system performs them, or why they form part of the optimal policy. So neither Dretske's conditions nor instrumentality alone distinguishes the system trained by supervised learning from the reinforcement learning agent which it mimics.

What does distinguish these two systems is that in reinforcement learning, the learning and reasoning that combine to determine the system's policy are themselves sensitive to instrumental relationships. This sensitivity plays a role in the development (and thus, later, the causal history) of these systems, and thus contributes to explaining their actions. In the supervised learning case the learning process is insensitive to such relationships, which explain their actions only in so far as they play a role in the origin of the training data. One way to describe the difference is that in the supervised learning case talk of instrumental value would merely be an interpretative gloss on the meaning of the target function, while in the reinforcement learning case sensitivity to this value is built into the algorithm.

4.4 *Large language models*

I now turn to a final example, which is Transformer-based large language models such as GPT-3 (Brown et al. 2020) and PaLM (Chowdhery et al. 2022). The basic form of these systems is as "foundation models" for language (Bommasani et al. 2021), which are trained on large quantities of data to predict the next word from a given sequence. This can be described as "self-supervised" learning because the data does not need to be labeled by humans. However, it is very like the supervised learning discussed so far. The system trains itself by generating a prediction for the next word, then observing the actual next word and using the difference to calculate weight updates. So the feedback that drives learning specifies the correct output for the previous input. Furthermore, whether the system samples inputs at random from a corpus or works its way through systematically, in the course of training its outputs do not affect its inputs.

Foundation models trained in this way on enough data, using the Transformer network architecture, are capable of producing remarkably fluent language and performing challenging linguistic tasks (Brown et al. 2020, Chowdhery et al. 2022). Their capabilities are

sometimes further enhanced by various forms of fine-tuning, including by reinforcement learning. For example, InstructGPT (Ouyang et al. 2022) is based on GPT-3 but fine-tuned by reinforcement learning for generic good performance in responding to prompts, as judged by human users.

Foundation models are not agents because they do not learn to produce outputs for their instrumental value. In training their outputs do not affect their future inputs, so it is impossible for them to learn to exploit such effects. This point is obscured by the way in which foundation models are often used, which is to extend prompts by many more words, so as to generate texts of dozens or hundreds of words. When they are used in this way, foundation models' outputs are immediately added to their inputs, so this is a situation in which agent-like capabilities could be useful. But a language-producing system cannot produce individual outputs for the sake of facilitating subsequent outputs unless it has been subject to training in which its outputs affected subsequent inputs, and unless it has a way to evaluate sequences of outputs.

A complication to this picture is that some language models, such as those used for sentence-to-sentence translation, use an algorithm called "beam search" (Sutskever et al. 2014). One way in which a translation system might work would be to select words to output one by one, based only on their probabilities conditional on the input and on previous words. However, it is intuitive that such a system would be outperformed by one which internally generated a sample of complete sentences and compared their relative probabilities, before committing to any output. This is what beam search involves: starting from a small number of likely first words, the algorithm explores branches of the trees of possible sentences that begin with those words. Beam search is not sufficient for agency, however, because in the translation case the outputs of the system are whole sentences, and they are not selected for their effects on future inputs. It may be possible for foundation models to learn to do something like beam search in the course of selecting their outputs—to select words partly by looking at which words could follow them—but even this would not be agency if it was done solely as a means to maximising the likelihood of the next word, as opposed to influencing subsequent inputs.

Although they are not agents, foundation models are noteworthy because Transformers seem especially well-suited to learning to predict the next item in a sequence. This means that they can be used to learn to model environments and to predict the consequences of their actions. This is not sufficient for agency, but it is a crucial step along one route to agency—the model-based method for selecting actions for their instrumental value. For example, consider a hypothetical chatbot based on a foundation model trained on human dialogue. This chatbot might be good at predicting how a human user would respond to some output, and thus how that output would affect the state of the conversation, making new outputs and subsequent responses possible.

Its predictive capacity would enable it to take instrumental actions, provided that it could also evaluate possible future conversation states and combine these abilities in action selection.

5. *Selection, functions and goals*

So far in this paper I have focused on descriptive differences between reinforcement learners and supervised learners. I have proposed that only reinforcement learners perform actions, because only their outputs are the result of processes which are sensitive to instrumental value. However, agency can also be seen—as I suggested in the introduction—as a species of norm-governed activity (again, understanding norms merely as non-arbitrary standards of success or correctness). A potential advantage of my account of agency is that the differences in history which matter for agency could ground normative differences. This is the idea which I will develop in this section.

The idea that an entity's history can give rise to norms to which its activities are subject is exemplified by the selected-effects theory of biological function (Garson 2016). This theory, which is a mainstream view in the philosophy of biology, claims that if a component of some organism exists because it was selected for a certain activity, the function of the component is to perform that activity. This means that the activities of the component are subject to a norm; the component may either function correctly or malfunction (or perhaps it may function better or worse, according to a standard derived from its selective history). Building on this claim, and following other authors, I will argue that learning, as well as selection, can give rise to norms governing the activities of the entities which these processes modify. I will then propose that processes of learning or selection can give rise to different kinds of norms. As well as grounding the functions of components or traits, such processes can also give rise to goals, which entail norms governing the activities of whole systems.

The central idea of the selected-effects theory is that functions arise from "consequence etiology" (Shea 2018). In natural selection, traits with effects which contribute to greater reproductive success tend to persist and proliferate in populations, while those with other effects tend to die out. This means that natural selection is one context in which we can explain why traits exist by citing their effects—or, more precisely, the effects of prior tokens of their type—and therefore a context in which a form of teleological explanation is consistent with naturalism. Learning is like selection in this respect, because it involves the persistence of phenomena which have effects of the right kind. Training neural networks involves preserving those combinations of weights which have the right effects, and modifying those which have the wrong effects; and reinforcement learning involves only repeating those actions which contribute, through their effects, to greater cumulative reward.

However, not all situations in which something exists or persists as a result of its effects seem to give rise to functions. This was roughly the theory of function proposed by Wright (1973), and many apparent counterexamples have been proposed. For example, a leak in a gas hose may persist because the gas poisons anyone who tries to repair it (Boorse 1976).

Rather than attempting to defend a more restrictive general theory of functions, Shea (2018) argues for a disjunctive account. He claims that natural selection and learning from feedback are both ways in which a feature can come to persist (be "stabilised") in a population or system in virtue of its effects, which are such that the feature will then have the function of bringing about those effects.[7] His rationale for including learning is that, like natural selection, it is a means by which complex systems are developed and modulated in nature which make it possible for organisms to bring about outcomes robustly, especially by using representations. Shea's project is to justify appeals to representation in explanations in cognitive science, and he claims that this is justified by the frequency with which we observe a certain abstract pattern: apparently-representational features are stabilised by natural selection and learning in the service of producing outcomes robustly.

This paper is not concerned with representations and focuses on non-biological learning. However, it remains true that learning from feedback, like natural selection, is a form of consequence etiology which can give rise to complex and cumulative adaptations and which enables systems to produce outcomes robustly. Furthermore, learning is—in all real cases—itself a trait which has origins either in natural selection or in the design of artifacts by intelligent agents. Forms of learning themselves have functions. This should give us greater confidence in attributing norms in this context.

The analogy between natural selection and learning from feedback is not perfect, but to the extent that there is an analogy, these processes map onto one another in the following way. Natural selection acts on populations, while learning acts on "systems"—including human and animal minds and computer systems of various kinds. In natural selection, traits of organisms become more or less prevalent in populations, with some becoming near-universal for extended periods, while in learning features of systems such as behavioural dispositions or combinations of network weights are preserved or modified, with some stabilised. Stability in both cases is a consequence of stable features of the environment. Reproduction and persistence or modification are both determined by feedback. In natural selection, organisms bear combinations of traits, these traits have effects on the environment, and these effects determine how many offspring the organisms will produce, thus causing traits to become more or less prevalent in the population. In

[7] Shea also claims that contributions to the persistence of an organism can ground functions, but this is less relevant to the issue at hand.

learning, systems produce outputs, these prompt feedback from the environment, and this feedback determines which features of the system will persist or be modified. The state of the environment which faces a new generation in the case of natural selection is analogous to the input to a learning system, and the traits of that generation are analogous to the features that determine the system's output.

Norm-generating processes of selection and learning therefore have the following five elements: an entity with features which are preserved or modified (a population or system); an environment; inputs from the environment to the entity; outputs with effects on the environment (with input-output transitions being determined by the features); and feedback from the environment, which determines which features are preserved or modified. This account gives us an abstract framework within which functions and goals, and the processes which give rise to them, can be described.

Functions arise when features of the entity which is affected by selection or learning are stabilised. The function of a stabilised feature of an organism or system is to perform the activity, or bring about the effect, that caused it to be stabilised. The effects of features cause them to be stabilised when they contribute to bringing about the right kind of feedback.

In contrast, systems come to have goals only in much more specific circumstances. What is crucial is how feedback leads to persistence and modification. In reinforcement learning, feedback consists of both reward and the next input. The system stores information about relationships between inputs and subsequent feedback, and uses this information in determining how to modify its features. Furthermore, these modifications follow rules which, in most environments, make a particular kind of feedback (greater reward) more likely. When these elements are in place, it is not only possible to explain the existence of features of the system in terms of the effects of their type, but also to explain some of the system's outputs in terms of the contributions that they tend to make bringing about greater reward over episodes of interaction with the environment. This kind of explanation involves attributing goals to whole systems, because it is whole systems which interact with environments across episodes, by producing sequences of outputs. Systems with goals also have features with functions, but entities with functional features do not always have goals, because they are not all formed by processes which respond to feedback in this specific way.

This account of goals is intended to be equivalent to my account of agency; all and only agents have goals in this sense. The systems with goals are those that perform actions, because actions are outputs that have been selected for their contributions to greater cumulative reward over episodes of interaction.

To test my proposal it would make sense to examine how it applies to biological cases. If the proposal implied that most animals are agents

while most other organisms, populations and sub-organismic systems are not, this would be some evidence in its favour. If it had other implications this might be evidence against. However, for this purpose it would be important to bear in mind that the account of goals which I have just offered is not intended to describe what it is for a person to have a goal in mind when performing an action, or for an animal to behave in a goal-directed way (as opposed to habitually; Dolan & Dayan 2013). Talk of goals and goal-directedness is widespread and these terms are used in several ways. Instead, I have offered an account of goals which is intended to mark a distinction between the norms governing agency and those governing other forms of activity. This is just one of the ways in which human activities can have goals.

6. *Conclusion*

I have argued that to be an agent an entity must come to produce outputs for their instrumental value. For this to be the case, the agent's dispositions must arise from processes of learning or reasoning which are sensitive to instrumental value. That is, the modifications that arise in agents as a result of feedback from the environment must be modulated by information about relationships between outputs, inputs and subsequent reward. One source of support for this account comes from the idea that agents characteristically pursue goals. This means that an agent's individual actions must be subject to standards of success according to their conduciveness to the agent's goals. The existence of such norms could be explained by the operation of learning and reasoning processes of the kind just described.[8]

References

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

Brafman, R. and Tennenholtz, M. 2002. "R-Max: A general polynomial time algorithm for near-optimal reinforcement learning." *Journal of Machine Learning Research* 3: 213–231.

Bommasani, R. et al. 2022. "On the opportunities and risks of foundation models." *arXiv* preprint.

Boorse, C. 1976. "Wright on functions." *Philosophical Review* 85: 70–86.

Brown, T. et al. 2020. "Language models are few-shot learners." *arXiv* preprint.

Buckner, C. "Deep learning: A philosophical introduction." *Philosophy Compass* 14 (10).

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

[8] Acknowledgements: I would like to thank Robert Long, Steve Petersen, Brad Saad, Derek Shiller and Jonathan Simon, and the participants at the Kathy Wilkes Memorial Conference, for their help with this paper.

Chowdhery, A. 2022. "PaLM: Scaling language modeling with Pathways." *arXiv* preprint.

Dennett, D. 1991. "Ways of establishing harmony". In McLaughlin (ed.). *Dretske and His Critics*. Oxford: Blackwell.

Dolan, R. and 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*. Cambridge: Bradford Books.

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.

Garson, J. 2016. *What Biological Functions Are and Why They Matter*. Cambridge: Cambridge University Press.

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

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

Lycan, W. 1987. *Consciousness*. Cambridge: The MIT Press.

McKenna, M. 2016. "A modest historical theory of moral responsibility." *The Journal of Ethics* 20: 83–105.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. Cambridge: The MIT Press.

Millikan, R. G. 1996. "On swampkinds." *Mind and Language* 11 (1): 103–117.

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

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

Papineau, D. 1993. *Philosophical Naturalism*. Oxford: Blackwell.

Russell, S. and Norvig, P. 2010. *Artificial Intelligence: A Modern Approach* (3rd edition). London: Pearson.

Sober, E. 1985. "Panglossian functionalism and the philosophy of mind." *Synthese* 64: 165–193.

Sutton, R and Barto, A. 2018. *Reinforcement Learning: An Introduction* (2nd edition). Cambridge: The MIT Press.

Ouyang, L. et al. 2022. "Training language models to follow instructions with human feedback." *arXiv* preprint.

Wright, L. 1973. "Functions." *Philosophical Review* 82: 139–168.

Zimmerman, D. 2003. "That was then, this is now: Personal history v. psychological structure in compatibilist theories of autonomy." *Noûs* 37 (4): 638–671.