Philosophy of Science (2026), **00**, 1–10
doi:10.1017/psa.2026.10261

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

Agency and Imitation

Patrick Butlin

Global Priorities Institute, University of Oxford, Oxford, UK
Email: patrick.butlin@gmail.com

(Received 11 April 2025 UTC; revised 07 July 2026 UTC; accepted 10 July 2026 UTC)

Abstract

AI provides a rich source of examples for theories of agency because it allows relevant features to be relatively freely combined. This paper considers agency in sequence models: models that have been trained to predict the next item in a sequence. Training models on sequences of observations and actions can yield systems that use reinforcement learning in context, as well as ones that merely imitate agents. Investigating systems from this group suggests a requirement that agents must learn about the environment in part from their own interaction with it. It also suggests a distinction between two approaches to understanding agency.

1. Introduction

What is an agent? What distinguishes agents from non-agents? AI researchers have long engaged with these questions (Wooldridge and Jennings 1995) and AI research has been defined as 'the study of the design of intelligent agents' (Poole et al. 1998, quoted in Russell and Norvig 2022). Accordingly, theories of agency have recently been proposed in response to developments in AI (Kenton et al. 2023; Butlin 2024, 2026). AI is also a rich source of challenging examples for theorists of agency because AI systems can have relatively unconstrained combinations of features. In this paper, rather than setting out to develop, defend or test a theory, I will consider some challenging examples—a group of systems that have been studied in recent AI research—to see how they might inform our understanding of agency.

The systems I will consider are *sequence models*: models trained to predict the next item in a sequence. Such models can be trained on sequences of observations and actions, allowing them to predict the most likely next action given an observation, some knowledge of prior observations and actions, and perhaps some further information. This capacity to select actions in response to observations means that sequence models can interact with environments and, if suitably trained, successfully perform tasks through this interaction. Large language models (LLMs) are sequence models, and they illustrate the capacity to interact with environments, since they can power compelling chatbots. LLMs also illustrate that sequence models can generate

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

new coherent sequences, rather than merely reproducing ones they have seen in training. Furthermore, sequence models can be trained and prompted to act in ways that tend to bring about specified goals, as I will explain below. This means that there is a case for thinking of some sequence models as agents.

However, one simple way to train a sequence model to bring about a goal through interaction with an environment is to train it to imitate a capable agent pursuing that goal. *Behaviour cloning* is a method in which standard supervised learning is used to train a model to imitate the behavioural trajectories of an 'expert' agent (or a collection of experts). While genuine agents may choose to imitate others, as humans often do, a system's being disposed to learn solely by imitation means that it is not sensitive to the consequences of its outputs, which could be seen as implying that it does not pursue any goal. Arguably, imitation is antithetical to agency. This line of argument suggests that some sequence models that are superficially agent-like are in fact not agents.

I will argue that the category of sequence models trained on sequences of observations and actions includes some models that are agents, some that are not, and some difficult intermediate cases. *Decision Transformer* (Chen et al. 2021) is a particularly challenging case and will be the focus of my discussion. I will suggest two ways of dealing with this case: first, an argument that Decision Transformer is not an agent; and second, a more conciliatory response, that distinguishes two ways of thinking about agency that suggest different views on the case. The second approach follows Wong (2025) in arguing that there are multiple conceptions of agency among researchers that should be disentangled.

Like Wong, other researchers have recently argued that the concept of an agent breaks down under scrutiny. Dung (2025) argues that the concept refers to a cluster of related but ultimately independent properties, and Abel et al. (2025) argue that agency is 'frame-dependent', in that classifying entities as agents depends on 'arbitrary commitments'. It seems to me, though, that the hypothesis that 'agent' is a useful, objective, binary category has not yet been disproven. Part of the motivation for this paper is to further test this hypothesis.

2. Clear cases of agency and non-agency

In this section, I will describe two possible AI systems and argue that the first is an agent while the second is not. These two systems are not sequence models; my aim in describing them is to establish precedents that will be useful when I turn to sequence models in subsequent sections.

As an example of an agent, consider a neural network trained by model-free reinforcement learning (RL) to play the video game Pong, using pixel maps as inputs (more concretely, this could be Mnih et al.'s (2015) deep Q-network). This system may be able to interact effectively with the Pong environment so as to achieve a high score. But beyond this, there are two deeper reasons to think of it as an agent.

First, model-free RL works in a way that justifies explanations of the system's behaviour that describe it as pursuing a goal.[1] In a value-based model-free RL algorithm such as Q-learning, systems maintain representations of the expected value

[1] For a similar argument, see Dretske (1988); for more detail, see Butlin (2022).

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

of performing actions in given states of the environment. These representations are initialised randomly, then updated in the following way: when the system performs an action in a given state, the value for that state–action pair is updated in the direction of the difference between it and the reward received plus the current estimate of the value of the best action in the new state. This means that actions that lead directly to high reward come to be represented as valuable. But what's more, it has the same effect on actions that lead *indirectly* to high reward. This is because, once the system has enough experience, it will represent that valuable actions—that might eventually lead to reward—are available from the states to which these actions lead.

When systems of this kind come to select actions, they typically choose those that they represent as having the highest values (although exceptions to this rule are needed to promote exploration). As we have just seen, they come to represent these actions as having high values because the actions are instrumentally conducive to states that yield high reward. So these systems work in such a way that when a type of action is conducive to a goal, this can cause them to take actions of the type—and most of their actions are caused in this way. This is excellent grounds to claim that these systems produce outputs *in order to bring about goals*.

Second, model-free RL appears to play a central role in human and animal agency. An extensive research programme uses the conceptual and computational framework of RL to explain how we learn and select actions (Niv 2009; Dolan and Dayan 2013). So the Pong-playing RL system works in a fundamentally similar way to paradigm agents. From now on, I will call it 'the Pong agent'.

This is not to say that the Pong agent is *uncontroversially* an agent. Some researchers who approach agency from a biological perspective would deny that this system is an agent on the grounds that it lacks autopoietic autonomy (Barandiaran et al. 2009; Virenque and Mossio 2024). But there are strong enough grounds for attributing agency to the Pong agent that it is reasonable to take this as a starting point for examining further examples.

As an example of a non-agent, consider a neural network trained by supervised learning to imitate the behaviour of the Pong agent—this could be similar to the systems trained to imitate human players using behaviour cloning by Kanervisto et al. (2020). Like the Pong agent, this network could take pixel maps as inputs and generate instructions to perform actions—to move the Pong paddle up or down—as outputs. However, it would be trained differently. Where the Pong agent generates its own training data by interacting with the environment and is updated using an RL algorithm that takes the game score as reward, this network would be trained to discriminate the states that are paired with different actions in a dataset generated by recording the Pong agent's behaviour. It would not see the game score or the consequences of the 'actions' it chooses in training. Nonetheless, it is possible that this system could also interact with the Pong environment so as to achieve a high score.

This system cannot be an agent because it is too similar to discriminative models trained to classify images, which are paradigm examples of artifacts that have functions but do not pursue goals. The system in question is trained in the same way as the image classifier, by supervised learning on a dataset of input–output pairs. Furthermore, its task can be conceived as one of image classification; it classifies pixel maps according to the different actions the Pong agent takes in response—moving

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

the paddle up or down, or leaving it in the same position. And it is likely to learn a fairly simple algorithm for doing this, since Pong is a simple game. The rule could be to always move the paddle up when the ball is above it and down when the ball is below it. We can imagine that in some circumstances the best way to accurately imitate expert agents is to understand the environment oneself and adopt and pursue the experts' goals; if there are cases like this, training for imitation could conceivably promote agent-like cognition. But the present case is not of this kind.

It may be useful for some purposes to describe all AI systems that interact with environments in ongoing ways as 'agents' (Sumers et al. 2023). But if agents are distinguished by how they learn and select actions, then, again, there are strong enough grounds to deny that this system is an agent that we can reasonably start from this claim.

3. Return-conditioned sequence models

As I have described, sequence models are models trained to predict the next item in a sequence. They are often based on the transformer architecture. Sequence models can be trained on sequences of interleaved observations and actions, allowing them to interact with environments in which those observations and actions are possible. The systems I want to focus on are *return-conditioned* sequence models, which have extra capacities because their training data also include *returns-to-go*.

Suppose we have records of a number of trajectories through an environment, including data about the rewards accrued at each time-step. These could be random walks or generated by an agent. Then we can organise this data so that each trajectory is made up of triples of the form [returns-to-go, observation, action], where 'returns-to-go' is the sum of the rewards subsequently accrued on that trajectory. We can now train a sequence model on this data and prompt it with our desired returns; the model will then tend to choose the actions that are most likely given that these returns will be achieved. Decision Transformer works in this way and was found to perform capably on a range of tasks including some Atari games (Chen et al. 2021).

A remarkable extension to this method is *Algorithm Distillation* (Laskin et al. 2022), in which trajectories collected from RL agents are used for training. These trajectories show improvements in the agents' policies in accordance with their learning algorithms. When Laskin et al. trained a transformer with a sufficiently long context window on this data they found that it learned an RL algorithm, which allowed it to subsequently improve its own performance through interaction with the environment.

Algorithm Distillation models are agents because they learn and use algorithms for in-context RL. Using these algorithms, these models learn in the course of each forward pass which types of actions are rewarding and select outputs accordingly. Although they are trained to imitate other agents, they learn to do so by responding to reward in a similar way.

Whether Decision Transformer is an agent is much less clear. In favour of agency, one might argue that in training this model learns about the structure of the environment—how states, actions and rewards are related—and that this information allows it to select rewarding actions in deployment, when suitably prompted. It can derive a policy that allows it to navigate the environment

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

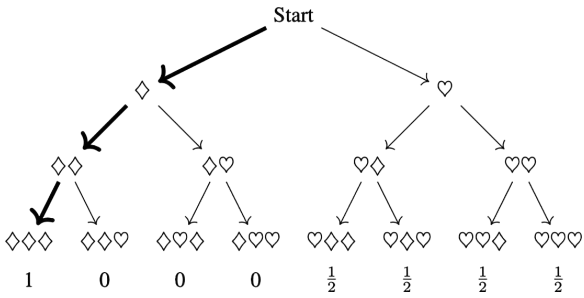


Figure 1. Three cards game. From Douglas et al. (2024).

successfully, achieving goals, from data about the environment that does not specify such a policy. But against agency, the same kind of argument that applied to the Pong-imitator system also applies to Decision Transformer. It is trained offline to predict the next token in a sequence, like an LLM or a behaviour-cloning model, and in this respect is like paradigm non-agents. Furthermore, in contrast with Algorithm Distillation, it does not learn to predict the next token *by* a recognisably agentic method.

4. The case against Decision Transformer

Because LLMs are sequence models and LLM-based agents are perceived to be central to the future of AI, researchers have considered the potential for sequence model agency. This work has identified characteristic failure modes to which sequence models such as Decision Transformer are susceptible when used as agents; these cast doubt on their status as agents.

The first failure mode is that sequence models are subject to *auto-suggestive delusions* (Ortega et al. 2021). For example, consider a sequence model trained on the behaviour of an expert stock trader, which does not include all of the information they use to make decisions. Suppose that the trader's practice is to call their broker only for one purpose: to make trades of over a million dollars. Now a situation can arise in which the sequence model is highly uncertain about what to do, because how the trader would behave depends on information to which it does not have access. In this situation, it might call the broker—perhaps the context implies that this is the most likely of several possible actions. Having called the broker, however, the model will assign a very high probability that the next action will be a trade with a value of over a million dollars, so this is what it will do. It will see its own previous output, of calling the broker, as very strong evidence that this is what happens next. That is, it will treat its own action as providing evidence that it lacked when it selected that action.

The second failure mode is *predictor-policy incoherence* (Douglas et al. 2024). This problem can arise because the training data will typically not reflect the model's own capabilities, with the consequence that the action most likely to succeed in the training data may not be the best action for the model. For example, consider Douglas et al.'s 'three cards game' (figure 1).

In this game, the player chooses three cards from a pack, taking either a diamond or a heart each time. If they choose three diamonds they receive a guaranteed

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

payment of one unit. If they choose any other set starting with a diamond they receive nothing, but if they choose any set starting with a heart they receive one unit with 50% probability. Now consider a return-conditioned sequence model trained on random walks through the three cards game, which is prompted to obtain a return of one unit. If this model chose a diamond at the first step, it would be guaranteed to succeed, because after this choosing further diamonds is the only action that ever leads to reward. However, this is not what the model will do. It will begin with the action that most often leads to a return of one unit in the training data, which is choosing a heart. Because the data is made up of random walks, beginning with a heart leads to reward 50% of the time, whereas beginning with a diamond only leads to reward 25% of the time. Despite having all the relevant information, the model fails to choose the option that guarantees success.

One possible interpretation of these weaknesses is that sequence models fail to distinguish between their own actions and other events. In the case of auto-suggestive delusions, sequence models behave as though they can learn about what to do from their own earlier actions. In the case of predictor-policy incoherence, sequence models behave as though their own subsequent actions will be similar to those in the training data, rather than being under their control. Ortega et al. argue that auto-suggestive delusions arise because sequence models fail to treat their own actions as uncaused causal interventions; the problem, they write, is that 'one cannot learn from one's own actions, but only from their effects' (2021, 6).

This line of thought seems to adopt a characteristic commitment of causal decision theory: that one should think of one's own actions differently from other events. In Newcomb's problem, causalists may say that what one should choose (two-boxing) is different from what one should want to happen (one-boxing, and thus receiving the higher reward). Something similar can be said of the three-card game: the action the agent should choose is different from the one that appears to succeed most often. It may be that some causalists believe that one is not really an agent if one does not think this way.[2]

However, this analysis is either over-intellectualised, in attributing mistaken reasoning to sequence models, or relies too heavily on talk of models behaving *as though* they make certain reasoning errors. It is unclear what treating one's own actions as different from other events means outside the context of reflective deliberation.

A better analysis is indicated by Douglas et al.'s (2024) finding that both problems can be alleviated by fine-tuning sequence models on their own behaviour. This reduces auto-suggestive delusions because in self-generated data the agent's actions no longer provide evidence that was not otherwise available, and reduces predictor-policy incoherence because in this data the observed actions are the agent's own. Furthermore, this is how simple model-free RL agents trained online avoid these two failure modes. They learn from data about the environment that they generate through their own interactions, and thus learn to select the actions that work for them. In contrast, sequence models imitate the behaviour of others that have different knowledge or dispositions.

[2] The decision theorist Timothy L. Williamson tells me that many causalists believe this, although to my knowledge this claim has not been made in the literature.

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

We could decisively resolve the question of whether Decision Transformer is an agent by adopting this as a criterion for agency. To be an agent, a system must learn about the environment in part through interacting with it. This condition enjoys intuitive support from the idea that a system that is set up to behave in a particular way prior to contact with the environment is not an agent. Agential autonomy, perhaps, consists in one's actions depending on one's own contact with the environment (Russell and Norvig 2022). On this view Decision Transformer would not qualify.

That said, objections to this view are possible. One possible objection is that we can imagine a system that interacts with an environment, using model-based planning to select actions, but is equipped with the model in advance. This system would not meet the proposed condition but would select actions in a way that is characteristic of agency (Butlin 2024).

5. Purist and pragmatic conceptions of agency

It is not obvious whether Decision Transformer is an agent. In favour of the view that it is an agent, one could argue that it learns about the environment and the distribution of rewards, then uses this knowledge to select rewarding actions, performing capably on some tasks. Against, one could argue that it learns wholly offline with a prediction objective, like a mere imitator, and that the fact that it doesn't learn from its own interaction with the environment leads to characteristic failure modes. As an alternative to trying to settle this dispute, we could attempt to diagnose it: to identify a tension between ideas about agency that explains why this case is difficult.

A promising way to do this may be to distinguish between a conception of agency on which agents work in a characteristic way, and one on which agents solve a particular kind of problem, perhaps imperfectly.[3] We can call the former a *purist* conception of agency and the latter a *pragmatic* conception. In the present case, I am imagining that the purist conception would deny that Decision Transformer is an agent while the pragmatic conception would accept it. But that need not be the case. There are a wide range of possible purist theories of agency and a wide range of possible pragmatic theories.

For a purist, being an agent is a matter of using agentic methods for learning and action selection. Characterising these methods is the central objective of research on agency, assuming a purist perspective. The claim that agents must learn from their own interactions with an environment could form part of a purist theory. One would expect the conditions in a purist account of agency to be related to standards of practical rationality—to be an agent is to learn and make choices as agents should—although agency must be consistent with some imperfections.

The pragmatic conception, in contrast, sees agency as a matter of functioning to solve a particular kind of problem, and thus as a matter of having certain capabilities. An initial account might describe the problem that agents solve as achieving goals

[3] As Sutton and Barto put it in the canonical textbook on RL, 'Reinforcement learning … is simultaneously a problem, a class of solution methods that work well on the problem, and the field that studies this problem and its solution methods' (2018, 2). My suggestion is that agents could be either those things that solve a certain problem, or those that use a certain class of solution methods.

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

through interaction with an environment. Agents vary in the generality, flexibility and reliability with which they are able to do this, and in the sources of their limitations. But goal-directedness and interaction are plausible necessary conditions for agency that may be seen as features of the task rather than the system. Pragmatic theorists of agency may well wish to add further conditions to this account, in order to narrow down the range of tasks seen as characteristically agentic—since, for example, the task that thermostats perform involves achieving a goal through interaction. For pragmatic theorists, developing and defending an account of the capabilities required for agency may be the central research objective.

Thinking of Decision Transformer as an agent accords with the pragmatic conception because the researchers who created it saw themselves as studying a new kind of solution to an agency-related problem. They write that they 'seek to study if generative trajectory modeling – i.e. modeling the joint distribution of the sequence of states, actions, and rewards – can serve as a *replacement* for conventional RL algorithms' (Chen et al. 2021, 3, emphasis in original). They go on to describe potential practical advantages of sequence modelling over RL. However, there are versions of the pragmatic conception that would entail that Decision Transformer is not an agent. The requirement that agents learn about the environment in part through interacting with it could just as well feature in an account of the problem that agents solve as in a purist theory describing a class of algorithms.

Hong Yu Wong (2025) distinguishes between architectural and behavioural perspectives on agency. According to Wong, the architectural perspective is the default position of most action researchers, and involves 'extrapolating and abstracting from details of the capacities and internal architecture of [paradigm agents] to arrive at accounts of agency and action' (4). Meanwhile, the behavioural perspective 'focuses on the interpretability of candidate systems as agents based on observing the system's behaviour' (5). One difference from my distinction is that I think of purists as focused on algorithms rather than architectures—but Wong also talks about 'mechanisms', which can cover both. The main difference, however, is that the pragmatic conception of agency need not be purely behavioural. The problem of agency can plausibly be seen as the problem of achieving behavioural performance on the basis of training data or methods that meet certain constraints. For instance, one could argue that behaviour cloning is not sufficient for agency because it requires possession of a satisfactory policy in advance, whereas return-conditioned sequence modelling may be sufficient, because it can be used to discover satisfactory policies. This is not a behavioural difference but a difference in the data on which the system is trained before it begins to interact with the environment.

6. Conclusion

The Decision Transformer and related models teach us two lessons about agency. One is that it is possible to solve agency-related problems in surprising ways, such as by sequence modelling. The successes of Decision Transformer and Algorithm Distillation are each striking in different ways; they remind us that in AI there are often many methods to achieve given levels of performance. This point means that performance—understood as suitability to solve certain kinds of problems—can come apart from algorithmic features, generating tension between purist and

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

pragmatist conceptions of agency. The second is that when systems are not trained on their own behaviour this can lead to characteristic failure modes. These failures provide additional support for the intuitive principle that agents must learn about the environment in part through interacting with it themselves.

Acknowledgements. Thanks to Rosa Cao and Nicholas Shea for arranging the symposium at which this paper was presented, and for discussion. Thanks to Timothy L. Williamson for discussion of causal decision theory.

Funding statement. None to declare.

Declarations. None to declare.

References

Abel, David, André Barreto, Michael Bowling, Will Dabney, Shi Dong, Steven Hansen, Anna Harutyunyan, Khimya Khetarpal, Clare Lyle, Razvan Pascanu, Georgios Piliouras, Doina Precup, Jonathan Richens, Mark Rowland, Tom Schaul, and Satinder Singh. 2025. "Agency Is Frame-Dependent." Preprint, arXiv:2502.04403. https://doi.org/10.48550/arXiv.2502.04403

Barandiaran, Xabier, Ezequiel Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Asymmetry, Normativity and Spatio-temporality in Action." *Adaptive Behavior* 17 (5):367–86. https://doi.org/10.1177/1059712309343819

Butlin, Patrick. 2022. "Machine Learning, Functions and Goals." *Croatian Journal of Philosophy* 22 (66):351–70. https://doi.org/10.52685/cjp.22.66.5

Butlin, Patrick. 2024. "Reinforcement Learning and Artificial Agency." *Mind & Language* 39 (1): 22–38. https://doi.org/10.1111/mila.12458

Butlin, Patrick. 2026. "The Agency in Language Agents." *Inquiry* 69 (6):2838–58. https://doi.org/10.1080/0020174X.2024.2439995

Chen, Lili, Kevin Lu, Arvind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. "Decision Transformer: Reinforcement Learning via Sequence Modeling." *Advances in Neural Information Processing Systems* 34: 15084–97.

Dolan, Ray, and Peter Dayan. 2013. "Goals and Habits in the Brain." *Neuron* 80 (2):312–25. https://doi.org/10.1016/j.neuron.2013.09.007

Douglas, Raymond, Jacek Karwowski, Chan Bae, Andis Draguns, and Victoria Krakovna. 2024. "Limitations of Agents Simulated by Predictive Models." Preprint, arXiv: 2402.05829. https://doi.org/10.48550/arXiv.2402.05829

Dretske, Fred. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: The MIT Press.

Dung, Leonard. 2025. "Understanding Artificial Agency." *The Philosophical Quarterly* 75 (2):450–72. https://doi.org/10.1093/pq/pqae010

Kanervisto, Anssi, Joonas Pussinen, and Ville Hautamäki. 2020. "Benchmarking End-to-End Behavioural Cloning on Video Games." In *Proc. IEEE Conference on Games (CoG)* 2020, 558–65. https://doi.org/10.1109/CoG47356.2020.9231600

Kenton, Zachary, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. 2023. "Discovering Agents." *Artificial Intelligence* 322:103963. https://doi.org/10.1016/j.artint.2023.103963

Laskin, Michael, Luyu Wang. Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, D. J. Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Minh. 2022. "In-Context Reinforcement Learning with Algorithm Distillation." Preprint, arXiv:2210.14215. https://doi.org/10.48550/arXiv.2210.14215

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Darshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518:529–33. https://doi.org/10.1038/nature14236

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press

Niv, Yael. 2009. "Reinforcement Learning in the Brain." *Journal of Mathematical Psychology* 53 (3):139–54. https://doi.org/10.1016/j.jmp.2008.12.005

Ortega, Pedro, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. 2021. "Shaking the Foundations: Delusions in Sequence Models for Interaction and Control. Preprint, arXiv:2110.10819. https://doi.org/10.48550/arXiv.2110.10819

Poole, David, Randy Goebel, and Alan Mackworth. 1998. *Computational Intelligence: A Logical Approach*. Oxford: Oxford University Press.

Russell, Stuart, and Peter Norvig. 2022. *Artificial Intelligence: A Modern Approach*, 4th edn. London: Pearson.

Sumers, Theodore, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. 2023. "Cognitive Architectures for Language Agents." Preprint, arXiv:2309.02427. https://doi.org/10.48550/arXiv.2309.02427

Sutton, Richard, and Andrew Barto. 2018. *Reinforcement Learning: An Introduction*, 2nd edn. Cambridge, MA: The MIT Press.

Virenque, Louis, and Matteo Mossio. 2024. "What is Agency? A View from Autonomy Theory." *Biological Theory* 19:11–15. https://doi.org/10.1007/s13752-023-00441-5

Wong, Hong Yu. 2025. "Interrogating Artificial Agency." *Frontiers in Psychology* 15:1449320. https://doi.org/10.3389/fpsyg.2024.1449320

Wooldridge, Michael, and Nicholas Jennings. 1995. "Intelligent Agents: Theory and Practice." *The Knowledge Engineering Review* 10 (2):115–52. https://doi.org/10.1017/S0269888900008122

Cite this article: Butlin, Patrick. 2026. "Agency and Imitation." *Philosophy of Science*. https://doi.org/10.1017/psa.2026.10261

https://doi.org/10.1017/psa.2026.10261 Published online by Cambridge University Press