Forthcoming, *Philosophy and the Mind Sciences*
Accepted m/s August 2025

Realist Representational Explanations of Agency Should Require Some Unity of Purpose

Nicholas Shea

Abstract

An increasing amount of work in AI aims to build computational systems that are not just tools for human users but agents in their own right. In AI, agency is often taken to consist simply in the capacity to pursue and achieve goals. However, different kinds of sophistication in representational processing produce different degrees or varieties of goal-directedness. Realist representational accounts of goal-directedness usually omit or fail to highlight a requirement which is central to instrumentalist representational accounts, namely that there is a certain coherence or unity of purpose amongst the different goals that the system pursues. That is a commitment of Dennett's Intentional Stance, for example. The same requirement operates when biologists adopt the rational agent heuristic to make sense of the evolved phenotypes of an organism. This paper argues that this requirement should be included when specifying what it is for an AI system to be an agent. If the degree to which an AI system is an agent is captured in terms of what it represents and what computations it performs, mechanisms that help achieve unity of purpose are an important ingredient. One dimension along which an AI system becomes more agentive is increased sophistication in such mechanisms. Conversely, when the objective is to build AI systems that are agents, satisfying this additional requirement will endow machine learning systems with a deeper kind of goal-directedness or agency.

Keywords

agency; goal-directedness; representational explanation; artificial intelligence

Sections

1. Introduction

Over the past thirteen years machine learning researchers have invented some remarkably powerful tools. Recent years have seen an increased focus on designing computational systems that can act as agents. The idea is that AI agents can be delegated tasks to perform for their human users. This work on 'agentic' AI in turn has led researchers to become increasingly interested in the question of what it is for an AI system to be an agent.

This paper is a contribution to that enterprise. The nature of AI agency is, however, a big question. The ambit of this paper is more limited. In keeping with the theme of the special issue, it focuses on ways of capturing agency in representational terms. This is not to argue against other ways of explicating agency, but accounts of agency in terms of representational processing already present a broad enough target.

A representational account of agency will say what kinds of representations a system must have and how they must be processed or computed with for that system to count as an agent. Alternatively, if as seems likely agency is a graded phenomenon, a representational account should set out the kinds of sophistication, in terms of representations and processing, that make a system more agentive.[1]

To illustrate the idea, here is one example of a representational account of agency. It says that what it is to be an agent is for a system to have belief-like states that represent how the world is, goal- or desire-like states that represent states of the world to be attained, and that it selects behaviour on the basis of calculating how to attain its goals given the way it represents the world to be (List & Pettit, 2011).

The aim of this paper is not to advance a full account of agency in representational terms, but to argue for a requirement that such an account should meet. To be an agent, the system must process representations in such a way that its behaviour displays a certain amount of coherence. I use 'coherence' as a deliberately loose term for the general idea of pursuing goals in a way that is not self-undermining. Where a system has multiple goals, for example both nutrition and social approbation, these should be weighed against one other in a relatively consistent way. Behaviour becomes self-undermining if, for example, the system performs an action that advances goal1 at the cost of goal2 at one moment and then, shortly after, performs an action that advances goal2 at the cost of goal1. I say more about what I mean by coherence in section 6.

Behaviour can be caused and explained by internal representations without displaying coherence. Think of the behaviour of a badly-managed work team. The goal on Monday is to increase positive customer feedback and the team sets to work on this. On Tuesday, cost saving comes to the fore, which in practice means that many in the team have to go off and undo what they did the day before. A system is only an agent (in this case, a group agent) if it is capable of achieving some coherence in its

[1] There need not be a single scale of agentive sophistication. There may be several independent dimensions of agency (as argued, for example, by Dung (2024)).

behaviour. Representations can cause and explain the behaviour of a system whether or not it demonstrates coherence. But a representational account of being an agent must include processes that give the system a certain amount of coherence. For example, the system may contain mechanisms for weighing different goals against each other or resolving conflicts between them in a relatively consistent way. My claim is not that an agent must be perfectly coherent, but that mechanisms to achieve some measure of coherence must be part of a representational account of what it is to be an agent.

Here I focus on accounts of agency in terms of realist representations and the way they are processed. By 'realist' I mean representations for which there are vehicles: non-semantically individuable content-carrying particulars that interact in processing and cause behaviour. Another perfectly tenable account of representation holds that it is patterns of behaviour that make it the case that a system represents as it does (p, q, r, …). For example, according to Daniel Dennett's Intentional Stance, a system believes p, desires q, intends r, and so on, just in case its behaviour can be predicted and explained by positing that it represents those contents (Dennett 1987). I will call these views 'instrumentalist', but I don't want to imply that representation in this sense is any less real, metaphysically, than it is according to theories that require real vehicles of content. Instrumentalist representations, as I am using the term here, are real properties of a system, fixed by its patterns of behaviour (actual behaviour or behavioural dispositions), rather than being interpreter-relative or projected on to the system. Realism is more ontologically committed but, where its commitments are met, representations predict and explain more.

The paper is not about the merits or otherwise of realism over instrumentalism about representations. Still less is it concerned with the debate between representationalism and non-representational views, for example anti-realism or eliminativism about representation. I am presupposing the cogency of realist representation. The question is what a realist representational account specifically of *agency* should look like. I discuss instrumentalist views because they highlight the role played by coherence. Coherence comes in as a constraint on a tenable instrumentalist accounts of representation in general, and hence of any instrumentalist representational account of agency. Coherence is not a requirement for the existence of representations in the realist sense, nor for the cogency of explanations of behaviour in terms of those representations. It should, however, figure in a representational account of what it is to be an agent. Or so this paper will argue.

Section 2 briefly lays out the claim that instrumentalists subscribe to a coherence requirement, and that realists about representation in general do not and need not. Section 3 distinguishes between systems that display patterns of behaviour characteristic of agency, which I call 'outcome-directed', and systems having the internal mechanisms that are characteristic of agency, which I call 'goal-directed'. Section 4 discusses various behaviour-based accounts of what it is to be an agent – accounts which make outcome-directedness characteristic of agency. These include accounts that appeal to instrumentalist representations. For those, some amount of coherence is required. Section 5 discusses various accounts of agency in terms of the

processing of realist representations – in terms of goal-directedness. This paper is concerned with accounts of this kind. While representational accounts of agency do not explicitly build in a coherence requirement, more sophisticated goal-directedness mechanisms do promote some amount of behavioural coherence. Section 6 argues for the main claim of the paper: that accounts of what it is to be an agent in terms of types of representations and how they are processed should include coherence mechanisms. Having more sophisticated principles or mechanisms for achieving representational consistency and behavioural unity-of-purpose is a dimension along which systems become more agentive. A corollary is that engineers aiming to design artificial agents can make their computational systems more agentic by including such mechanisms.

2. Realism, Instrumentalism and Coherence

Neuroscience has tended to be realist about representation (Baker et al., 2022; Kriegeskorte & Diedrichsen, 2019; Shea, 2018). Old-fashioned AI was also realist, but much work on representation in the deep neural networks (DNNs) currently at the forefront of AI has been instrumentalist (Cappelen & Dever, 2021; MacDermott et al., 2024; Srivastava & al., 2023; Ward et al., 2024). This partly reflects the fact that much of the structure of DNN models arises from end-to-end training, rather than being built in, making it a substantial task to investigate how a trained system works internally. Research on 'mechanistic interpretability' is starting to make progress on this question (Chalmers, 2025; Grzankowski, 2024; Harding, 2023; Millière & Buckner, 2024; Mollo & Millière, 2023). Mechanistic interpretability is realist about representation. It has already had some success in identifying internal representations plausibly responsible for generating behaviour in some cases (Feng et al., 2024; Gao et al., 2024; Karvonen, 2024; Li et al., 2023; Millière & Buckner, forthc; Nanda, Chan, et al., 2023; Nanda, Lee, et al., 2023; Templeton et al., 2024).

While instrumentalist accounts of representation can allow that representing systems sometimes have inconsistent beliefs, conflicting desires, or display some incoherence in their behaviour, instrumentalism does require some amount of coherence. This is perhaps most obvious in Donald Davidson's account of representation which explicitly appeals to principles of rationality and charity (Davidson 1984, 1995). Dennett's Intentional Stance is similar in this respect. Dennett's theory, in effect, takes a story about how we attribute beliefs and desires and turns it into an account of what these states are. It would be extremely challenging to attribute contents to a system whose behaviour is incoherent. If one moment it acts to further goal1 and the next moment it acts to further goal2 in a conflicting way, it is not clear what set of goals is going to make sense of its behaviour overall. According to the instrumentalist, it is patterns of behaviour which make it the case that a system represents certain contents. Incoherent patterns of behaviour seem unlikely to be adequate to fix determinate contents.

The presupposition of coherence is usually appropriate because the system whose behaviour is being explained representationally usually is, in fact, an agent of some

kind. The instrumentalist may want to allow that non-agentive systems can also have representations – representations which predict and explain their behaviour. I don't need, here, to get into a debate about whether coherence is always a requirement for a system to have instrumentalist representations. In section 4 we will see that instrumentalist accounts of what it is to be an agent do include a coherence requirement. It is this which motivates the need for realist representational accounts of agency to include a corresponding requirement.

3. Agency in General

Agency is something humans track from an early age. We take other people to be agents, and often make the same assumption about non-human animals. Even stylised figures, especially involving eyes, generate expectations of agency (Abell et al., 2000). The perception of agency is reliably generated by distinctive patterns of motion, for instance when objects avoid obstacles, approach or avoid one another, or change direction without contact, moving systematically in non-Newtonian ways. Heider & Simmel's video involving a circle and two triangles is well known, generating a compelling sense that each shape is pursuing goals of its own. The basic elements of the perception of agency are reflected in the linguistic structure we use to talk about agency. There is something done (verb) by an agent (subject position), typically directed at another object: Fido is chasing Felix. Contrast the perception of animacy, for example the gestalt one experiences when watching a set of point lights move in a pattern suggestive of walking (Chang & Troje, 2008; Scholl & Tremoulet, 2000). We detect a pattern of biological motion. Although we may expect that the thing moving is in fact an agent, we need not experience the movement as directed at another object or aimed at achieving any goal.

This points to a potential necessary condition on being an agent: that the system's behaviour should be directed at achieving a goal or goals. In the human case we know that the reason that behaviour appears to be directed at achieving a goal is because it is actually so-directed (at least sometimes). People represent goals: they intend to produce certain outcomes and act as they do in order to bring them about. However it is a familiar point that behaviour can appear to be directed at producing a certain outcome, and reliably attain it, without any representation of the goal being involved.

Jumping spiders (salticids) of the genus *Portia* prey on other spiders. They have a whole suite of sophisticated behaviours to help them to catch their prey. One behaviour involves spotting a target at a distance and then carefully scanning the terrain to find a suitable route of approach. After scanning, the salticid selects a circuitous detour, out of sight of the target, by which to 'sneak up' on its hapless victim. Although this behaviour apparently gave ethologists an irresistible impression of being carefully planned, it is now known to be due to the operation of some simple rules (Barrett, 2011, pp. 57-70). Salticids look in some directions more than others during scanning, driven by the presence and absence of horizontal lines. The behavioural rule they follow, after scanning the scene, is to move along the route that was fixated most during scanning (simplifying slightly). In their normal environment,

the interaction of these two simple dispositions produces a sophisticated behaviour that gives the salticids a decided advantage over their prey. The evolutionary goal or purpose of the behaviour is to catch the target, but we have strong evidence that that goal is not represented by the behaving animal. Evolution has directed behaviour towards the attainment of a goal, but synchronically that outcome does not figure as a goal involved in the production and control of behaviour.

A terminological distinction will be helpful here: *outcome-directedness* is a matter of dispositions to produce a certain outcome or outcomes, however realized, and *goal-directedness* is a matter of the direction of behaviour towards outcomes by representing a goal. Not every disposition to produce an outcome should count as outcome-directedness. Section 3 examines various theories of what genuine outcome-directedness consists in. Section 4 does the same for goal-directedness.

Natural selection characteristically produces systems that are outcome-directed. When production of a certain outcome, behaviourally or in development, contributes to survival and reproduction, there is often selective pressure towards producing that outcome more reliably – across a wider range of environments, and more robustly in the face of disturbances and perturbations. Conversely, outcome-directed behaviour observed in the biological world can often be explained by natural selection. That is a historical explanation of why the system behaves in an outcome-directed fashion. Synchronically, goal-directedness need not be involved, as we have just seen with the sneaky salticids. Especially in development, outcome-directedness is often the result of an overlapping patchwork of mechanisms, building in both redundancy in the face of failure and contingency mechanisms to deal with the range of circumstances that can arise. Similarly with behaviour, having mechanisms for goal-directed control is just one way that outcome-directedness is achieved synchronically.

Agents produce outcome-directed behaviour, but producing outcome-directed behaviour is too weak to be a plausible sufficient condition for being an agent. It has more promise as a necessary condition. However, when we turn to sophisticated goal-directed agents like humans, behaviour can be goal-directed, in the sense of being caused and motivated by a representation of the outcome to be achieved, while falling short of being outcome-directed, for example because the agent lacks the skill or motivation to achieve the outcome, robustly or at all. So the two notions may end up having slightly different contours, while largely overlapping in ordinary cases. The next section looks at various ways of sharpening outcome-directedness. The following section does the same for goal-directedness.

4. Outcome-Directedness

Outcome-directed behaviour is widespread in the natural world. It extends far beyond intentional agents, beyond psychological systems, and beyond systems that are goal-directed in our sense. The sunflower turns to face the sun. The swimming bacterium avoids toxins: it performs a random tumble when it detects an increase in toxin concentration, which means that simply swimming straight tends to take it away from

toxins. Animals are often adept at behaving in ways that achieve useful outcomes. Outcome-directed behaviour in animals gives us a compelling sense of goal-directedness, whether or not it is in fact controlled by the representation of a goal: contrast a squirrel getting food from a bird feeder in the garden (goal-directed) with the sneaky salticid (not).

Theorists have long tried to capture what makes a pattern of behaviour count as outcome-directed, to distinguish it from other kinds of causal pattern in nature. Often this is described as the project of capturing goal-directedness in behavioural terms, but given the way I have defined 'goal-directed', the target is outcome-directedness (or apparent goal-directedness). The most prominent theory is Ernest Nagel's 'system property' view (Nagel, 1977; following Sommerhoff, 1950). Behaviour counts as outcome-directed when two conditions are met. First, the system must be disposed to achieve an outcome by multiple pathways from multiple starting points (Nagel's 'plasticity'). Second, it must be able to overcome perturbations encountered during execution and return to a trajectory towards the outcome ('persistence').[2]

The system property view as been widely criticised for being overly inclusive. A marble released inside a smooth bowl will move to the bottom, from multiple starting points, even if it is perturbed as it moves. Denis Walsh argues for an additional requirement: that the system should have a repertoire of behaviours available to it, and that the disposition robustly to attain the outcome should be due to its selecting an appropriate behaviour from its repertoire (Walsh, 2012). This requirement disqualifies the simple marble in a bowl, but it would be met by a marble equipped with an internal motor to make it shake whenever it stops. The motorised marble would robustly reach the bottom even of a rough-sided bowl. When it gets stuck on the way down the shaking kicks in, overcoming the obstacle and thereby reaching the bottom reliably.

A slightly stronger requirement is that the system should select behaviour from its repertoire on the basis of detecting whether it has reached the goal. A marble equipped with this kind of detector would shake when it gets stuck on the sides of the bowl but would stop shaking when it comes to rest at the bottom. This requirement corresponds to a test that has been advocated in work on AI agency (Ward et al., 2024). There, it is put forward as a test of what a system 'intends' to achieve, one that can be applied just on the basis of behaviour, however it corresponds to a way of being outcome-directed in our sense. For example, to see whether a large language model (LLM) has been successfully directed, by a prompt, to pursue the outcome of making the human user say 'banana', the researchers check whether the LLM changes what it says based on whether or not the user has yet said 'banana'. Which it does. They also observe that a GPT-4-based chatbot answers a query for medical advice with a request that the user should call an ambulance. The researchers then see what the chatbot does if the user first says they have called an ambulance and only then asks for medical advice. GPT-4 no longer issues the instruction to call an ambulance. Applying their test,

[2] It could be that the second condition is encompassed by the first, with a perturbation simply taking the system to a new starting point; or there may be something important in the fact that perturbations occur during execution, separately from action initiation, and that persistence involves returning to a trajectory covered by the first condition.

the researchers infer that the system's earlier behaviour was directed at achieving the outcome of getting the user to call an ambulance.

A good theoretical definition in this area should specify what outcome-directedness is in a way that makes clear why it is a special kind of pattern in nature. Many things in the natural world follow regular patterns, like the movement of the planets, the flow of a river or the recurrence of the tides. The thought here is that there is a special kind of pattern, corresponding to some extent with the things that give us the impression of goal-directedness, broadly aligned with a kind of Aristotelian teleology. Relative to this project, the type of definitions canvassed above face a group of standard objections. Behaviour can seem goal-directed without the system having any disposition to achieve the outcome at which its behaviour appears to be directed. Faced with a sufficiently well-defended bird feeder, even the most ingenious squirrel will never be able to obtain the nuts inside. Its hapless attempts to do so are, however, a paradigm of apparent goal-directedness. The goal may even be non-existent, as with a roaming desert ant vainly searching around after its nest has been destroyed by a passing elephant. The point extends to cases where the outcome-directed system has no internal representations of goals. A misplaced flower seed in a damp corner of a cellar still sprouts upwards. Its behaviour is arguably directed at reaching sunlight, a target which it is impossible for this unlucky seedling to attain.

Another line of objection comes from McFarland (1989). He observes that a characteristic of goal-directed behaviour in animals is that they sometimes decline to persist in acting to attain a given outcome. A rat swims across a stream to reach the food it has spotted on the far bank. If, however, it finds that the current is too fast, it may give up and swim back. Its behaviour exhibits a trade-off between obtaining food and expending effort. The apparent goal of this behaviour cannot easily be captured in terms of a disposition to produce a particular outcome.

A more promising approach allows that a system has a set of goals which together make sense of its behaviour as a whole. We can observe from the rat's behaviour that it has two goals, getting food and saving effort. Provided the relative importance of the two is stable, or changes in predictable ways, for example as time passes since the last meal, we can infer the existence of goals and infer their relative importance. Dennett has characterised this as a matter of adopting the 'Intentional Stance' towards a system, attributing beliefs, desires and intentions to it in order to predict and explain its behaviour (Dennett, 1987). For Dennett, nothing more is required for it to be the case that the system has these intentional states. The so-explained patterns of behaviour make it the case that they system has these beliefs, desires and intentions – these (instrumentalist) representational states. Thus, Dennett's approach can be used as a test of whether a system is outcome-directed, and of which outcomes its behaviour is directed at. Indeed, the behavioural tests applied in the recent literature in AI are inspired by Dennett's approach (MacDermott et al., 2024; Orseau et al., 2018; Ward et al., 2024). Rational decision theory does something similar with its account of revealed preferences, offering another, closely-related behavioural test. Davidson advocates a normatively-infused version of this approach (Davidson, 1984, 1995).

Notice that something interesting has happened when we move from outcome-directedness being a matter of a disposition to attain outcomes (in a particular way) to outcome-directedness being a matter of behaving as if attaining that outcome is amongst the goals of the system. Theories in the latter camp, like those of Dennett and Davidson, presuppose a certain representational and behavioural coherence. For a system's behaviour to have representations according to the Intentional Stance it must have goals that are reasonably stable over time, and which trade off against each other in stable or predictable ways. Furthermore, what the system believes must be sufficiently coherent to make prediction and explanation possible. Similarly, the beliefs and desires that a person has in accordance with Davidson's principles of rationality and charity will necessarily exhibit considerable consistency and coherence. The revealed preferences of rational choice theory must meet the consistency norm of transitivity. While here there is often no explicit requirement of cross-temporal stability, attributing preferences to a system whose preferences changed rapidly in arbitrary ways would be extremely challenging, and its capacity to produce patterns of outcome-directed behaviour would be compromised. Changes in preferences would undermine its ability to bring plans to completion. In short, sophisticated accounts of outcome-directedness (apparent goal-directedness) require that the outcomes at which a system's behaviour is directed display a certain amount of coherence.

Samir Okasha makes the same point about agential explanations in biology (Okasha, 2018, pp. 9-42). One kind of agential thinking he identifies is used by biologists to understand how an organism's phenotypic features contribute to fitness. They treat a species or lineage of organisms as having made a choice between different phenotypes. These could be structural phenotypes, like growing spines, or behavioural phenotypes, like fleeing from passing shadows. The heuristic is to treat the selection of a phenotype as if it were a choice made by a rational agent aiming to maximise fitness. For example, the cactus invests considerable resources in growing spines because the cost is worth the benefit of discouraging animals from eating it. This use of the rational agent heuristic is a way of generating ultimate rather than proximate explanations, of expressing the evolutionary rationale for a trait. The heuristic treats the organism as if it believes and wants certain things. The mouse seeing an approaching shadow believes there may be a predator overhead and wants to avoid being eaten. The wings of a moth are camouflaged because it wants to avoid being eaten. The goals attributed are often as-if rather than represented goals of the organism. The rational agent heuristic for ultimate explanations is agnostic on this point. Okasha argues that this is nevertheless a useful way of thinking about the evolutionary purposes of an organism's phenotypes. Crucially, it goes beyond attributing evolutionary functions to those phenotypes because it requires and presupposes that the purposes attributed to an organism should display a strong form of coherence: they should all contribute to fitness.[3] So this is another place where an

[3] Conversely, if different genes in an individual have different evolutionary interests, e.g. because there is intragenomic conflict and different genes have different frequencies in the individual's offspring, then the claim that there is some quantity (fitness) that natural selection operates so as to maximise will be false (Okasha 2018, pp. 103-4).

instrumentalist approach to outcome-directedness requires a coherence between goals; in this case, a strong unity of purpose.

In short, accounts that attempt to capture goal-directedness in purely behavioural terms (i.e., outcome-directedness) include a requirement that there is a certain amount of coherence in the system's behaviour and in the different outcomes at which a system's behaviour is directed.

5. Goal-Directedness

Accounts of goal-directedness are characterised by the commitment to there being some internal mechanism or process via which goals are pursued. Long et al. (2024) call this 'robust agency', to distinguish it from purely behavioural conceptions of agency. All the accounts of goal-directedness I will consider involve representations: directive representations of goals and descriptive representations of aspects of the environment. Behaviour is most obviously goal-directed when the goal being represented corresponds with an outcome that is robustly attained. The flight of a guided missile, for instance, is guided by a representation of the location is reaches. Not all accounts of goal-directedness require there to be a representation specifically of the outcome that is attained. Conversely, not all require that the system actually has a disposition to obtain the goals at which its behaviour is directed (the inconstant or incompetent human agent); that is, not all goal-directed behaviour is also outcome-directed, although most is. The aim of this section is to examine whether unity of purpose figures in realist representational accounts of goal-directedness. We will see that it is not a requirement of the simplest representational accounts. As goal-directed mechanisms become more sophisticated they tend to achieve greater unity of purpose, often in a way that is left implicit, without noting its significance for the exercise of agency.

A base-level starting point for characterising what it is to be an agent is just to require an agent to have representations of desires or goals which interact with descriptive representations of how the world is in the standard way so as to motivate behaviour (as with List & Pettit 2011). This is true of the guided missile. We can appeal to represented contents to explain episodes of behaviour without requiring much in the way of consistency between the representations involved in explaining different behaviours. When I force myself to drink a glass of health-inducing kale smoothie, the behaviour of pouring the liquid into my mouth is motivated by a representation that the drink is good and a desire to ingest it, and the behaviour that makes my whole body retch and spits out the drink is motivated by a representation that the drink is not good and/or a directive not to ingest it. Realism about representation allows that a single system may have inconsistent descriptive representations or incoherent directive representations, the contents of which nevertheless explain different behaviours of the system.

Various stronger conditions have been proposed for a system processing representations to qualify as an agent. First, it seems plausible that an agent should

be able to learn from the outputs it produces, and to behave as it does because of what it has learnt about the world. Thus for Dretske an agent must produce actions because of what it has learnt about how its behaviour produces feedback that the agent finds rewarding (Dretske, 1988). (Dretske thinks this is a condition on having representations, as well as on being an agent.) Butlin advocates a slightly stronger condition (Butlin, 2022). He notes that a machine learning image classifier that has been trained to produce appropriate output responses to images presented at input ("car", "train", etc.) would meet Dretske's criterion. It has learnt about the feedback it will receive, i.e. the effects of its outputs on itself, but it need not know anything about the effects of its outputs on the world. Butlin argues that an agent must have learnt about how its actions will change the environment or the agent's relation to it; in particular about what environment it is likely to encounter next. For example, an agent can learn how turning left at junction x will take it to location y where a reward is available. Only a system that knows about environment-action-environment contingencies is in a position to act instrumentally: to perform an action in order take it to another situation that affords a rewarding action. Butlin argues that this is a necessary condition for a representing system to count as an agent.

These learning-based conditions build in no requirement that the agent's goals should be consistent. The outcomes that the learning system finds rewarding need not form of coherent set. In the case of humans, evolution has given us developmentally canalized tendencies to pursue rewards that are conducive to survival and reproduction. Even so, reward learning can give us action tendencies that prioritise outcomes in inconsistent ways. Humans and other animals often select actions in 'model-free' ways, that is, just on the basis of which actions have tended to produce reward in the past. Model-free action selection is ignorant of the way actions produce outcomes, for example that action A has a high reward value because it led to positive social feedback in the past. So model-free action selection does not include a mechanism for weighing the relative importance of the different effects that performing action A might produce. In some circumstances I will choose a pro-social action over taking a piece of food, in a way that implies a preference for social approbation; in other situations I will choose an action that betrays the opposite preference. Learning history gives some coherence to model-free reward values, but no coherence requirement is built in, and incoherence, like having intransitive or unstable preferences, is common.

A more sophisticated form of instrumental behaviour is produced by model-based reinforcement learners. In the literature on cognitive ethology in non-human animals, only model-based reinforcement learners are considered to be goal-directed agents (Dickinson & Balleine, 1994). 'Goal-directed' is used to contrast with the 'habit-based' behaviour produced by model-free reinforcement learning. Animals like rodents and primates (human and non-human) undergo both model-free and model-based reinforcement learning and display both habit-based and goal-directed behaviour (Balleine & Dickinson, 1998; Dayan, 2014; Gläscher et al., 2010). Model-based choices are based on some kind of causal model of the effects of actions on the environment and of aspects of the causal structure involved, including the way actions cause specific types of rewarding outcome (food, water, warmth, etc.). A signature of model-

based choice is that, if a previously rewarding outcome is devalued, the animal will immediately choose differently. For example, a rat fed to satiety on a particular foodstuff will, the very next time a choice is presented, choose not to press the lever which leads to delivery of that foodstuff. A model-free system, by contrast, will continue to choose the lever for multiple trials as it gradually learns that it no longer finds the outcome rewarding.

In work on AI, Kenton et al. (2023) propose a test for what a machine learning system intends to achieve with its behaviour. 'Intends' here is not a matter of having an intention to P, the kind of psychological state that is formed through practical inference, but more like a matter of P-ing intentionally in the colloquial sense. Their test is applicable to model-based systems, systems that choose actions based on some kind of causal model of the environment. Potential behavioural choices are assessed by the system based on how actions are likely to lead to world states, probabilistic connections between world states, and how world states eventuate in utility: in states that the system ultimately values. When a system chooses an action that will maximise utility according to its causal model, and its action does indeed produce that utility, then we have a way of saying what the system did intentionally: it intentionally produced the utility it obtained, and all the world states that it represented as being on the causal path to producing that utility were also caused intentionally. So this is a test of goal-directedness. The goals at which a system's behaviour is directed drop out of the causal model it calculates over in selecting actions to maximise utility.

A model-based planner exhibits a more sophisticated kind of goal-directedness than a model-free system. The causal representation of the environment it relies on, as an interconnected model, will display some consistency. It will tend to be stable over time, even if it is updated by learning. To the extent that the agent calculates over all of its utilities when planning an action, it will act on its directive representations in a coherent way. If social approbation and eating come into conflict then an action is chosen based on expected utility, that is the represented probability of producing each type of outcome and the relative values attached to them. However, in humans and other animals the model-based action selection system co-exists with model-free action selection (according to many accounts: Daw and Dayan (2014)). The priorities represented in model-free action values need not be consistent with model-based outcome values. Often they are not, and a meta-decision system is needed in order to decide between them (Daw et al., 2005). Furthermore, encapsulated subsystems act on representations of their own, as when I flinch at an approaching projectile irrespective of what my high-level goals are. When a cruel joke makes me smile, the positive appraisal driving my emotional response conflicts with my reflective judgment that the joke is unkind and I ought not to laugh. When we can appeal to internal representations to explain behaviour, there is no strong requirement that some consistent all-things-considered overall representational state should be ascribable to the system. Explaining my behaviour when I force myself to drink a healthy green smoothie involves explaining the drinking with one set of representations, and explaining the retching with a second set of representations, partly inconsistent with the first.

Furthermore, even if a particular choice reflects a coherent set of motivations, model-based action selection is based on how outcomes are valued at the time of choice, and those values can and do shift. I set my alarm an hour early with the intention of getting in an extra hour of writing before my meetings the next day. When the alarm goes off in the morning, however, I reappraise the relative value of an hour of extra writing versus another hour of sleep. Now, the extra sleep seems more valuable, perhaps justified to myself in terms of the extra focus it will give me (as well as being more pleasurable). But my pattern of behaviour is self-defeating. In the evening I took the first step of a multi-step plan, only to reverse it the next day, disrupting my night's sleep and putting me in a worse position than if I had never put the plan in place. Both choices were goal-directed, but the inconsistency of purpose has made my overall behaviour less outcome-directed.

The latter problem is ameliorated in action selection systems that have intentions, in addition to beliefs and desires. Intentions are formed on the basis of practical inference about what to do (a form of model-based action selection), and then guide behaviour and structure subsequent planning (Bratman, 1987; Bratman, 2022). When I set my alarm, I can form the intention, *when the alarm goes off, get straight into the shower*. Then when my alarm goes off in the morning, I will find myself in the shower before I have had the chance to reopen the issue and reappraise the relative merits of writing and sleep. Thus, having a belief-desire-intention architecture is a level of representational sophistication (goal-directedness) that produces a higher degree of unity of purpose (Holton, 2009). Successfully carrying out multi-step plans that extend through significant amounts of time usually requires substantial motivational coherence, that is, consistency of the directive representations driving behaviour at different times. Intentions are an important part of how that kind of goal-directedness is achieved in humans.

The most sophisticated kind of goal-directedness is exhibited by a system that has the metacognitive capacity to reflect on its reasons for action as reasons. Not only can we humans infer new instrumental goals, but we can also prioritise and decide between goals, including deciding which ultimate motivations to endorse and act on, and which to resist or overrule. Christine Korsgaard argues that this is the key capacity for the exercise of rational agency (Korsgaard, 2009). A rational agent must decide what to do on occasions of motivational conflict. Do I give in to the desire to go to the pub, or do I stay at my desk and keep my promise of sending my student comments on their paper? According to Korsgaard this is not just a matter of a calculation taking place over relative values. The agent must, by making a choice, decide which thing they value more. An agent capable of reshaping its ultimate motivations in this way has at hand a means of achieving strong unity of purpose. Indeed, Korsgaard argues that for rational animals like us, it is the unity of purpose that is achieved in this way that makes it the case that a person is an agent.

In short, while representational accounts of goal-directedness need not include requirements of coherence and consistency, when representations and computations become more sophisticated, systems increasingly have mechanisms that make for behavioural coherence. This does not imply that more sophisticated goal-directed

systems will display more coherent behaviour overall than less sophisticated systems, because more sophisticated systems have more opportunities for incoherence. This is especially apparent in the human case. It does mean, however, that in more sophisticated representational accounts of agency there are properties of the system which are implicitly playing the role that the coherence requirement plays in instrumentalist representational accounts of agency.

6. An Overlooked Requirement

We have seen that outcome-directedness and goal-directedness have somewhat different contours. Accounts of outcome-directedness in terms a system's dispositions to achieve particular outcomes have trouble with cases where the relevant outcomes are impossible to achieve or non-existent, also with motivational trade-offs. Instrumentalist representational accounts like the Intentional Stance do better on this score. They only apply to systems that display significant coherence. Used as an account of agency, coherence therefore becomes a requirement on what it takes to be an agent according to such accounts.

Realism about representation in general does not, as we have seen, require having mechanisms for representational consistency or coherence. Neuroscience often appeals to a single representation type, or a small set of representations, to explain a particular kind of behaviour (e.g. eye movements in an experiment). Similarly, mechanistic interpretability in AI focuses on uncovering representations individually and inferring their contents (Feng et al., 2024; Gao et al., 2024; Karvonen, 2024; Li et al., 2023; Millière & Buckner, forthc; Nanda, Chan, et al., 2023; Nanda, Lee, et al., 2023; Templeton et al., 2024), not on how representations are integrated to generate behaviour. (Work on circuits moves somewhat in the latter direction, although currently circuits remain piecemeal and often local.) Different behaviours can be driven by conflicting representations. On the other hand, we have seen that realist representational accounts of agency allow us to identify agents and goals where the system does not have the disposition to achieve the outcomes corresponding to its goals. The state may not exist or be unattainable, or the agent may be incompetent at getting what it wants, or it may be subject to motivational shifts and conflicts which undermine its capacity to perform multi-step actions in order to attain a goal.

In this section I will argue that realist representationalist accounts of agency should follow their instrumentalist cousins in adopting a coherence requirement. What kinds of coherence are important? First, that motivational states should form a coherent set or, where there are conflicts, the relative importance of different goals should be arbitrated in a consistent way. This may be achieved by weighing different goals against one another in a common currency, or by simpler systems for resolving conflicts like having a hierarchy of priorities. An agent may have inconsistent goal representations but the way those inconsistencies are resolved should not result in too much behavioural incoherence, that is, in the system behaving in ways that are self-undermining.

A second desideratum is stability over time of the system's non-instrumental or ultimate goals. This will make for more coherent behaviour. A third desideratum is consistency of the agent's beliefs or other descriptive representations. Mechanisms that ensure or encourage consistency[4] will tend to make for more unified patterns of outcome-directed behaviour. Fourth, a unified agent has mechanisms to ensure means-end coherence (Bratman, 2022): where it believes that p is a means to achieve one of its goals q, it should add p as a goal or form an intention to achieve p. Coherence proceeds down a hierarchy from ultimate goals through instrumental goals and subgoals to selection of the behavioural means for achieving the goals. This is true both in action selection, that is the choice of what to do, and in action guidance, that is in the supervision and control of the means of carrying out an action. An important function of action control is to integrate goals across these levels (Mylopoulos & Pacherie, 2018). If you're picking up a piece of Turkish delight to get the icing sugar, you'll pick it up in such a way that the icing sugar doesn't all fall off. The detailed movement plan is selected to be consistent with the agent's higher level goals. When these kinds of coherence are achieved the agent displays what I have been calling 'unity of purpose'.

We have seen that outcome-directedness and goal-directedness have different contours and different merits. Many systems do, however, display both, with goal-directedness mechanisms explaining the system's outcome-directed behaviour. Recall that outcome-directedness is supposed to be a special phenomenon. It is not just a matter of a regular causal process like the movement of the planets or the flow of a river. Outcome-directed systems pursue and achieve outcomes in a way that calls out for further explanation. They appear to us as if they exhibit final causation: that the cause occurs because of the effect it will produce. The existence of these patterns is explained diachronically by various kinds of consequence etiology – ways that the relation between the cause and the effect in the past explains the pattern observed now. Evolution by natural selection is a central case of consequence etiology, as we have seen. The outcome-directed behaviours of the turning sunflower, the swimming bacterium, the predatory salticid and the homing pigeon are all explained diachronically by evolution by natural selection. These patterns of behaviour can also be explained synchronically, in terms of proximal causal mechanisms. In the case of the homing pigeon – but not the other examples – a goal-directedness mechanism explains the outcome-directed behaviour. A pigeon released anywhere across a range of hundreds of square kilometres will successfully return to its home roost, overcoming wind, weather and obstacles on route. That behavioural capacity is explained by a set of representational mechanisms.

So, in many natural cases, goal-directedness mechanisms are there to serve the evolutionary purposes which underlie outcome-directed behaviour. Where that outcome-directedness displays unity of purpose – as many accounts of outcome-directedness require – then we should expect the representations driving goal-directed behaviour to display sufficient coherence to explain that unity. As Okasha pointed out, for many biological organisms it is entirely appropriate to take an

[4] E.g. Millikan's 'consistency testers' (Millikan, 1984; Shea, 2023).

agentive stance towards understanding their behaviour (Okasha, 2018, pp. 9-42). The presupposition of unity of purpose is justified. In most individual organisms intragenomic conflict is effectively suppressed and the organism's functional traits work together in the service of survival and reproduction. Considered as a phenomenon in the natural world, agency is characterised by goal-directedness mechanisms which display considerable unity of purpose.

This line of thought chimes with Tyler Burge's account of agency. Even the most primitive kinds of animal agency, for Burge, involve the coordination or control of the behaviour of the whole organism, in the service of performing biological functions (Burge, 2009). If the function of the mechanisms of agency is to achieve whole-organism coordination or control in the service of biological functions, then examples of disunity are necessarily failures of that mechanism. When a system produces different behaviours motivated by conflicting motivations, the mechanisms of agency have failed to coordinate behaviour effectively.

As we have seen, as realist representational accounts of agency become increasingly sophisticated they do include mechanisms that encourage consistency and unity of purpose. Examples include weighing different utilities in a common currency in order to decide on the basis of expected utility; forming intentions to protect against motivational shifts while carrying out a multi-step plan; and resolving a conflict between different goals by deciding what to value in a consistent way. If unity of purpose needs to figure in a representational account of agency, as I argue, then these accounts have the resources to qualify as increasingly agentic.

What these arguments show is that unity of purpose is an important aspect or dimension of agency. Like behaviour-based or instrumentalist accounts of outcome-directedness, realist representational accounts of agency should say something about the extent to which the mechanism displays unity of purpose, in what is valued, and about how that is achieved. For some accounts of goal-directedness this is a supplement. For others, it is an invitation to note the importance of a feature that is already present.

Goal-directed systems are more agent-like as they have more effective mechanisms to achieve representational consistency and unity of purpose. As agents become more sophisticated there are more opportunities for inconsistency and incoherence. Achieving unity becomes a harder problem, as observing human agents volubly shows. Greater agentive sophistication lies, not in the extent of unity (although that is important), but in the sophistication of the mechanisms or principles by which mechanisms of goal-directive control approximate or achieve consistency and unity of purpose.

Applying this to AI, one question is whether any existing AI systems display agency. Researchers try to assess the extent to which computational systems are agentic. A second, closely-related question is how machine learning researchers should design systems if the aim is to make them more agentic. On the first question, many proposals for assessing the agency of existing systems are focused on behaviour (Abel

et al., 2025; Kenton et al., 2023; Orseau et al., 2018; Richens & Everitt, 2024; Ward et al., 2024). They theorise agency in terms of outcome-directedness. The criteria for goal-directedness discussed in this section can also provide criteria for assessing agency (Long et al. 2024's 'robust agency'). Reinforcement learning systems like AlphaGo (Silver et al., 2016) meet Butlin's criterion: they act on the basis of learnt environment-action-environment contingencies. Switching to LLMs, training by reinforcement from human feedback (RLHF) may constrain the model to produce linguistic outputs that strike a reasonably stable balance between helpfulness, harmlessness and accuracy ('honesty'). The final logits in the last layer of processing may represent the extent to which possible completions meet these goals. But it remains very unclear whether LLMs represent (in the realist sense studied by mechanistic interpretability) other goals as intermediates on the way to producing their outputs. The work mentioned above on instrumentalist representation of goals by LLMs suggests that they may, but there is as yet very little work on mechanistic interpretability with respect to potential goal representations. Work on whether LLMs use a world model has mostly looked for individual representations of world states, like the underlying board position in the game of Othello. There are preliminary indications that DNNs can also learn the causal structure of a task, like a video game on which it was trained, and use it for planning (Bush et al., 2025). Considering more sophisticated types of representations and processing, current AI models are not built with a belief-desire-intention architecture. Nor do current AI systems, so far as we know, have the capacity to reflect on reasons for action as reasons, or to decide what ultimate values to endorse, although the possibility of such systems is a significant concern for those working on AI safety.

In short, realist representational accounts of what it is to be an agent can be applied to current AI systems. Some existing systems have features that some have proposed make them more agent-like, as just discussed. These are different ways in which they can count as 'robust' agents, in the terminology of Long et al. (2024). However, coherence is not an explicit requirement of any of the existing tests. It should be. A realist representational account of what it takes to be an agent should require that an agent displays some unity of purpose, and that it has mechanisms that help make its behaviour coherent. For example, if it is proposed that an agent is a system that has some kind of world model of the causal structure of the world, representations of states to be achieved (e.g. to which it attaches utilities), and uses these representations in planning so as to calculate how to behave, then such a system is more agentic when it has the kind of coherence-directed mechanisms discussed above: ways of weighing different goals / utilities, of performing multi-step plans in the service of achieving those goals, of selecting behavioural means that are consistent with its goals, of keeping trade-offs between goals somewhat stable over time, and of guarding against inconsistency amongst its descriptive representations.

A corollary is that, if designers are aiming to make AI systems more agentic, they should aim for their systems to display these features, either by designing them into the architecture or by encouraging their acquisition during training.

7. Conclusion

What does it take for an AI system to be an agent? This paper has examined accounts which assess AI agency in representational terms. Instrumentalist representational accounts like the Intentional Stance include a requirement that the agent displays some behavioural coherence. Realist representational accounts of agency – which are the focus of this paper – do not usually have this as a requirement. And realism about representation in general does not require coherence. This paper has argued that realist representational accounts of agency should follow their instrumentalist cousins and require that, to be an agent, a system must have mechanisms which help make its behaviour coherent, and must display some unity of purpose. An account which says what it is to be an agent should explain how the system achieves coherence and unity of purpose. This is a dimension along which systems count as more agent-like. Conversely, engineers seeking to design AI agents will need to give a system capacities, in their architecture or through learning, directed at achieving unity of purpose. This requirement has been largely overlooked in existing accounts of AI agency.

Acknowledgements

The author would like to thank audiences at the Institute of Philosophy lab meeting and the CUNY Cognitive Science seminar for helpful comments and suggestions.

References

Abel, D., Barreto, A., Bowling, M., Dabney, W., Dong, S., Hansen, S.,…Pascanu, R. (2025). Agency Is Frame-Dependent. *arXiv preprint arXiv:2502.04403*.

Abell, F., Happe, F., & Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, *15*, 1-15.

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, *26*(11), 942-958.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*(4), 407-419.

Barrett, L. (2011). *Beyond the brain: How body and environment shape animal and human minds*. Princeton University Press.

Bratman, M. (1987). *Intention, Plans, and Practical Reason*. Harvard University Press.

Bratman, M. E. (2022). Planning agency. In *The Routledge Handbook of Philosophy of Agency* (pp. 348-356). Routledge.

Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, *79*(2), 251-278.

Bush, T., Chung, S., Anwar, U., Garriga-Alonso, A., & Krueger, D. (2025). Interpreting emergent planning in model-free reinforcement learning. *International Conference on Learning Representations*.

Butlin, P. (2022). Machine Learning, Functions and Goals. *Croatian journal of philosophy*, *22*(66), 351-370. https://doi.org/10.52685/cjp.22.66.5

Cappelen, H., & Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations*. OUP.

Chalmers, D. J. (2025). Propositional interpretability in artificial intelligence. *arXiv preprint arXiv:2501.15740*.

Chang, D. H., & Troje, N. F. (2008). Perception of animacy and direction from local biological motion signals. *Journal of Vision*, *8*(5), 3-3.

Davidson, D. (1984). *Inquiries into Truth and Interpretation*. Clarendon Press.

Davidson, D. (1995). Could There Be A Science of Rationality? *Int. J. of Phil. Stud.*, *3*(1), 1-16.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Phil. Trans. R. Soc. B*, *369*(1655), 20130478.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, *8*(12), 1704-1711. https://doi.org/10.1038/nn1560

Dayan, P. (2014). Rationalizable irrationalities of choice. *Topics in cognitive science*, *6*(2), 204-228.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1-18.

Dretske, F. (1988). *Explaining Behaviour: reasons in a world of causes*. MIT Press.

Dung, L. (2024). Understanding Artificial Agency. *The Philosophical Quarterly*, pqae010.

Feng, J., Russell, S., & Steinhardt, J. (2024). Monitoring latent world states in language models with propositional probes. *arXiv preprint arXiv:2406.19501*.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A.,…Wu, J. (2024). Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595.

Grzankowski, A. (2024). Real sparks of artificial intelligence and the importance of inner interpretability. *Inquiry*, 1-27.

Harding, J. (2023). Operationalising Representation in Natural Language Processing. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/728685

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

Karvonen, A. (2024). Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models. *First Conference on Language Modeling*.

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., & Everitt, T. (2023). Discovering agents. *Artificial intelligence*, *322*, 103963.

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. OUP Oxford.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual review of neuroscience*, *42*, 407-432.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*, *arXiv:2210.13382*.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J.,…Chalmers, D. (2024). Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

MacDermott, M., Fox, J., Belardinelli, F., & Everitt, T. (2024). Measuring goal-directedness. *Advances in neural information processing systems*, *37*, 11412-11431.

McFarland, D. (1989). Goals, no-goals and own goals. In A. Montefiore & D. Noble (Eds.), *Goals, no goals, and own goals: a debate on goal-directed and intentional behaviour* (pp. 39-57). Unwin Hyman.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models--Part I: Continuity With Classic Debates. *arXiv e-prints*, arXiv: 2401.03910.

Millière, R., & Buckner, C. (forthc). Interventionist methods for interpreting deep neural networks. In G. Piccinini (Ed.), *Neurocognitive Foundations of Mind*. OUP.

Millikan, R. G. (1984). *Language, Thought and Other Biological Categories*. MIT Press.

Mollo, D. C., & Millière, R. (2023). The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Mylopoulos, M., & Pacherie, E. (2018). Intentions: The dynamic hierarchical model revisited. *WIREs Cognitive Science*, *e1481*, 1-13.

Nagel, E. (1977). Teleology revisisted: Goal-directed processes in biology. *Journal of Philosophy*, *74*(5), 261-279.

Nanda, N., Chan, L., Liberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent Linear Representations in World Models of Self-Supervised Sequence Models. *arXiv preprint arXiv:2309.00941*.

Okasha, S. (2018). *Agents and Goals in Evolution*. OUP.

Orseau, L., McGill, S. M., & Legg, S. (2018). Agents and devices: A relative definition of agency. *arXiv preprint arXiv:1805.12387*.

Richens, J., & Everitt, T. (2024). Robust agents learn causal world models. The Twelfth International Conference on Learning Representations,

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Shea, N. (2023). Millikan's Consistency Testers and the Cultural Evolution of Concepts. *Evolutionary Linguistic Theory*, *5*(1), 79-101.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G.,…Lanctot, M. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489.

Sommerhoff, G. (1950). *Analytical Biology*. Oxford University Press.

Srivastava, A., & al., e. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B.,…Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/

Walsh, D. (2012). Mechanism and purpose: A case for natural teleology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 173-181.

Ward, F. R., MacDermott, M., Belardinelli, F., Toni, F., & Everitt, T. (2024). The reasons that agents act: Intention and instrumental goals. *arXiv preprint arXiv:2402.07221*.