Causal asymmetry from the perspective of a causal agent

Peter W. Evans[*1], **Gerard J. Milburn**[†2], **and Sally Shrapnel**[‡2]

[1] *School of Historical and Philosophical Inquiry*, University of Queensland
[2] *School of Mathematics and Physics*, University of Queensland

March 23, 2021

Abstract

Agency accounts of causation are often criticised as being unacceptably subjective or anthropocentric. According to such criticisms, if there were no human agents then there would be no causal relations, or, at the very least, if humans had been different then so too would causal relations. Here we describe a model of a causal agent that is not human with a view to exploring this latter claim. This model obeys the known laws of physics, and we claim that it endows the causal agent with a "*causal viewpoint*: a distinctive mix of knowledge, ignorance and practical ability that a creature must apparently exemplify, if it is to be capable of employing causal concepts" (Price, 2007, p.255). We argue that this model of a causal agent provides a clear illustration of the epistemic constraints that define such a 'causal perspective', and we employ the model to demonstrate how shared constraints lead to a shared perspective. Furthermore, we use this model to scrutinise the alignment of three familiar asymmetries with the causal asymmetry: the thermodynamic arrow, the arrow of time, and the arrow of deliberation and action.

1 Introduction

Agency accounts of causation are often criticised as being unacceptably subjective or anthropocentric. According to such criticisms, if there were no human agents then there would be no causal relations, or, at the very least, if humans had been different then so too would causal relations. The origins of agency accounts of causation stretch back to Hume (1748), whose elucidation of causation in terms of the habit forming behaviour of humans has a distinct psychological, agent-centred character (Beebee, 2007). Contemporary versions, however, are set apart from other manipulability accounts of causation by an emphasis on the role of the agent in the process of exploring causal structure in the world. Whereas some manipulability accounts, such as the interventionist accounts of causation of Pearl (2009) and Woodward (2003), consider causal relations to be an objective, external feature of the world, the most prominent current agency account interprets the notion of an intervention as inseparable with the ascription of causal relations to the world as a function of the agent perspective (Price, 2007; Ismael, 2016).

[*]email: p.evans@uq.edu.au
[†]email: g.milburn@uq.edu.au
[‡]email: s.shrapnel@uq.edu.au
All authors contributed equally to this work.

According to Woodward (2003, p.123), an agency account "leads us toward an undesirable kind of anthropomorphism or subjectivism regarding causation". But this raises a question concerning exactly how significant the role of the 'human' agent is to agency accounts. Could it be plausible that any suitably competent nonhuman agent could generate robust causal knowledge of the world (whether or not this causal knowledge be best understood as a projection on, or a discovery of, the world)? Here we explore this possibility through a detailed examination of the model of a causal agent proposed by Milburn and Shrapnel (2020), where the agent consists of specialised subsystems that comprise sensors, actuators, and a learning machine. While this modelled causal agent is not a human agent, the model is described by the known laws of physics, and we claim that it endows the causal agent with a "*causal viewpoint*: a distinctive mix of knowledge, ignorance and practical ability that a creature must apparently exemplify, if it is to be capable of employing causal concepts" (Price, 2007, p.255).

We claim that this model of a causal agent provides a notable counterpoint to the criticism that agency accounts are unacceptably subjective or anthropocentric. In particular, we argue that this model provides a clear illustration of the epistemic constraints that define a 'causal perspective'—a notion that we introduce in §2—and we employ the model to demonstrate how shared constraints lead to a shared perspective. The causal perspective that Price (2007) identifies for agents such us ourselves is a function of the deliberative orientation of the agent: which events the agent considers to be known or fixed, and which events the agent considers unknown and open to manipulation by the agent. Following Evans (2020), we note that the causal perspective is also a function of the specific physical capability of the agent: what *kind* of variables an agent identifies for the purposes of manipulation and causal modelling, and what pragmatic choice an agent makes to separate the world into system and environment.

We describe a physical model of a minimal causal agent in §3. We take this model to be an elegant illustration of the role of both deliberative orientation and physical capability in generating the agent's causal perspective, as well as promoting a better delineation of what is meant here by 'perspective'. In particular, we argue that the deliberative orientation of the agent is necessarily aligned with the thermodynamic directedness of its actuators, sensors, and learning machine. We connect this model to the debate regarding agency accounts in §4, where we demonstrate how the asymmetry in the causal relations ascribed by the agent to the world can arise as a result of the agent's own thermodynamic bias. Moreover we argue that this model of a causal agent illustrates a key feature of causal perspectivalism: that a class of agents sharing a causal perspective will agree on the ascription of causal relations to the world. Thus, far from agency accounts being unacceptably subjective or anthropocentric, we claim that the model shows that agency accounts are perfectly capable of characterising causation in the absence of human agents, and can achieve a kind of objectivity as a function of intersubjective agreement on causal ascriptions.

In addition, we use this model of a causal agent to scrutinise the alignment of three familiar asymmetries with the causal asymmetry: the thermodynamic arrow, the arrow of time, and the arrow of deliberation and action. We claim that the deliberative stance of the agent is a function of the necessarily irreversible dynamics of the sensors, actuators, and learning machine of the agent, and so is intimately tied to the thermodynamic gradient under which the agent labours.

2 Causal perspectivalism, anthropocentrism, and objectivity

There has been disagreement concerning whether agency accounts of causation are perniciously anthropocentric, and so not capable of providing an account of causation suitable for underpinning the objectivity of scientific inquiry (in so far as the goal of scientific inquiry is to uncover functional relations between variables characterising physical systems of interest). Causal per-

spectivalism has emerged as a key component of agency accounts, and has developed as a response to claims of anthropocentrism. Evans (2020) outlines this disagreement in detail, and we provide here briefly the main contours of that debate concerning anthropocentrism and objectivity in agency accounts, and how causal perspectivalism is placed to dissolve the disagreement.

The origin of causal perspectivalism stretches back to Menzies and Price (1993), who develop what is known as an agency account of causation, a subspecies of manipulability or interventionist accounts. According to such a view, while causation is to be understood in terms of manipulations on the world to bring about effects, an agency account makes the extra claim that our "ordinary notions of cause and effect have a direct and essential connection with our ability to intervene in the world as agents" (Menzies and Price, 1993, p.187). The substance to the view of Menzies and Price is that causation should be understood as a *secondary quality*, much the same as colour. In the case of colour, it is well established that our perceptual capabilities as human agents play a crucial role in defining colour perception. Likewise, they argue, the intervention capabilities of human agents play a crucial role in defining causal relations and, in particular, the asymmetry of causation. Crucial for our purposes here, a traditional criticism of the agency account (and one that Menzies and Price (1993, p.198–201) do in fact address) is that the account is too anthropocentric, in the sense that, if there were no human agents, then there would be no causal relations.

The main proponent of this criticism against the agency account is Woodward (2003) (see also Woodward (2007, 2009)): "it leads us toward an undesirable kind of anthropomorphism or subjectivism regarding causation" (Woodward, 2003, p.123). Woodward's own position is that causation is an 'objective' relation, and he employs his own interventionist account of causation to emphasise this point. The full detail of Woodward's account is not so important here, but it will be instructive to consider a brief outline. Consistent with other manipulability accounts, the interventionist account defines a causal relation formally in terms of the possibility of an intervention on some variable characterising a physical system that changes the probability distribution of the possible values of some other variable, holding fixed all other variables relevant to the system. Woodward then defines in detail what exactly is meant by an 'intervention' (Woodward, 2003, p.98). For our purposes here we can simply state that an intervention must satisfy a series of conditions that place the intervention in total control of the cause and eliminate any correlations between cause and effect that are not a function of the intervention.

A causal relation identified by an intervention is embodied in a functional relation between cause and effect, which must be invariant over some range of possible interventions, stable over some range of background conditions, and independent of other functional relations that might characterise other parts of the system in question. It is thus only within such an appropriate range for both interventions and background conditions where the functional relation can be established, and under the condition that different causal mechanisms are distinct, that we can utilise the functional relation by manipulating the cause as an appropriate means for manipulating the effect. For this reason, whether the relation between two variables is a causal relation is relative to a range of context dependent factors.

Thus, the identification of a causal relation boils down to being able to intervene on a system as per the conditions that formally constrain the intervention, and the possibility of finding appropriate ranges of interventions and background conditions under which sufficiently distinct functional relations hold. There is a strong sense in which the 'relativity' of the relevant functional relations to interventions and background conditions is related to a 'coarse-graining' of the variables and relations taken to characterise some system. As Evans (2020, p.10) puts it:

> Variables and functional relations with these properties may manifest themselves at finer or coarser grains. The appropriate level of grain at which to model a system is dependent upon

> the sort of causal information one wishes to obtain by way of intervention. Likewise, whether a system can be characterised at all as being constituted by causal relations will itself depend upon the particular coarse-graining that is chosen, and we coarse grain as part of the modelling process *just so* dynamical variables with the right sort of functional interrelationships can be objectified for our practical purposes.

Despite this, Woodward ultimately argues that, according to his interventionist account, causal relations are objective: "Relative to a specification of system and a level of description or graining for it... once one fixes the variables one is talking about, it is [an] 'objective' matter whether and how [the variables] are causally related" (Woodward, 2007, p.90). Underpinning his "objectivist position regarding the connection between causality and agency" is that "quite independently of our experience or perspective as agents, there is a certain kind of relationship with intrinsic features that we exploit or make use of when we bring about B by bringing about A" (Woodward, 2003, p.125).

Woodward is explicit in his criticism of the agency account of Menzies and Price that they "are not very forthcoming about just what is meant by their claims that causation is "projected" onto the world by us or "constituted" by our beliefs and attitudes" (Woodward, 2003, p.118). In response to this criticism, Price (2007) develops a more nuanced account of the role of the agent in an interventionist account. Key to this more nuanced account, which Price calls "causal perspectivalism", is to recognise that there are natural constraints generated by our own particular epistemic perspective when we encounter a system on which we wish to intervene. For human agents these natural constraints arise because we are "constructed, situated, embedded and oriented in time" (Price, 2007, p.252), and are constituted by our having knowledge of the past, but not of the future, and this limits the direction of the functional relations between variables that are exploitable by us by underpinning the distinctions we make between exogenous and endogenous variables. Thus, according to causal perspectivalism, the distinction between cause and effect is reduced to an agent's perspective: "the strong temporal asymmetry of the notion of intervention—and hence, apparently, of our causal thinking in general—stems not merely from the fact that we are agents, but also from a further contingency concerning our temporal circumstances: above all, the strong temporal bias of our epistemic access to our environment" (Price, 2007, p.280).

The 'perspective' here is defined by the 'situatedness' or 'embeddedness' of the agent. Taking our lead from sociological approaches to the philosophy of science, the concept of situatedness recognises that agents are necessarily embedded within a social context or environment (Anderson, 2019). In this case, however, the notion of environment is extended to a physical context: agents are necessarily physically embedded in an environment with a temporal orientation.[1]

Despite this perspectivalism, there is a sense in which the objectivity of the functional relations between variables characterising a system remains. Conditional upon the epistemic constraints specific to, say, human agents, there is a subsequent fact of the matter concerning which relations are causal and which are not (Evans, 2020, p.11). In the terms of Woodward's interventionist account, once we have specified a coarse-graining according to which the right interventionist conditions are met, the functional relations, and so the causal relations, are 'objective'. It is for this reason that Woodward has grounds to claim that causal relations according to his view are objective. But Price's elaboration of causal perspectivalism makes it clear exactly how this objectivity is dependent upon a particular perspective—as Price (2007, p.279) puts it, the interventions themselves become a "Trojan Horse" for causal objectivists. Since the human perspective is generated by inescapable epistemic constraints, this perspective is stable across human agents, which promotes strong intersubjective agreement between agents and thus a

[1] While Price consistently refers to the temporal orientation of the environment, we could equally well consider an agent that is embedded in a physical environment with a local thermodynamic gradient.

sense of 'objectivity'. Price (2017) revisits this point to conclude that the alleged 'objectivity' of Woodward's view and the alleged 'subjectivity' of agency accounts are really not such different accounts (see also Slezak (2009)).

As noted by Evans (2020, p.11), however, "the dependency of interventions on the agent perspective is not limited to the temporal bias of our epistemic access to our environment". This is most clearly seen in the coarse-graining that preconditions the possibility of exploiting a causal relation by way of intervention: the context dependent factors that allow exploitability are all perspectival, agent-dependent systematisations of the manipulable parts of the world. And the particular process of coarse-graining that an agent undertakes is inherently connected to the physical capabilities of the agent. As Evans (2020, p.11) puts it:

> Relative to a specification into system and environment, and a level of description or grain, there is a fact of the matter concerning what causes what. But this specification and level of grain are agent-centric features of the causal model of some system. Thus while Price diagnoses the objectivity of interventionist causation as residing in the fact that causal relations are indexed to a perspective defined by the temporal embedding of the intervening agent, we can see that there are other agent-centric pragmatic constraints, like the specification of a level-of-grain, that contribute to the perspective also.

This rapprochement between the objectivity of causal relations and the anthropocentrism of the agency account is made more precise by Ismael (2016). According to Ismael, the functional relations we exploit as causal agents are generated by an invariant "modal substructure" of reality, and we partition the invariant structure into cause and effect based on our idiosyncratic epistemic constraints and limitations concerning that structure. Thus, when an agent makes a pragmatic separation of the world into system and environment, and elicits a coarse-graining of the system into an appropriate set of variables as per the above interventionist conditions—both anthropocentric, perspective-dependent specifications—then there is a fact of the matter, established by the modal substructure, concerning whether some relation between variables is causal.

It is significant for our purposes here to note that agents that inescapably share idiosyncratic epistemic constraints, as human agents do, will share a 'perspective' when it comes to identifying causal relations, as we noted above. But these epistemic constraints are not only a function of a particular temporal orientation, as Price (2007) noted, but also a function of the particular physical capabilities of agents as manipulators and detectors of the world. Not only are agents embedded in an external physical environment, but they are also embedded in a more immediate environment defined by their own physical capabilities. These capabilities not only determine the kind of variables that an agent will identify for the purposes of manipulation and causal modelling, but also heavily influences the pragmatic choice an agent will make when separating the world into system and environment. Shared constraints amongst agents such as temporal orientation and physical capability ensure the stability of an agent perspective across a class of agents, and so define an equivalence class of agents that constitutes a 'perspective'.

3 Physical model of a causal agent

Recall that our goal in this work is to argue that the above epistemic constraints that define a causal perspective are illustrated by the model of a causal agent proposed in Milburn and Shrapnel (2020). In particular, we take this model to illustrate the role that the agent's deliberative orientation and physical constitution play in generating its causal perspective. We also claim that the model sheds light on the relationship between three familiar asymmetries that align

with the causal asymmetry, and key to this is showing that the deliberative stance of this agent is a function of the necessarily irreversible dynamics of its network of sensors, actuators, and learning machine, and so is intimately tied to the local thermodynamic gradient under which the agent labours.

While grounding causal perspectivalism in the physics of agents might seem hopelessly ambitious—in so far as the causal agents with which we are most familiar are complex biological systems—we are emboldened by recent efforts to construct artificial causal agents such as robots (Russell and Norvig, 2010) and embodied AI (Pfeifer and Bongard, 2006; Pfeifer and Gabriel, 2009). We take such artificial agents to be complex enough to begin to illustrate elements of agency accounts of causation and, in particular, causal perspectivalism. So let us turn to a minimal model of a causal agent in order to set out this illustration.

3.1 A minimal model of a causal agent

A minimal causal agent is an open system maintained in a non-thermal-equilibrium steady state which is stabilized by access to a low entropy source of energy. By 'open' here we mean that (i) the agent can exchange energy with the environment in the form of heat, and (ii) the agent has access to an external work reservoir or power source. The agent is constituted by a number of specialised subsystems including sensors, actuators, and a learning machine.[2] A pictorial representation of this minimal model is depicted in Fig. 1.

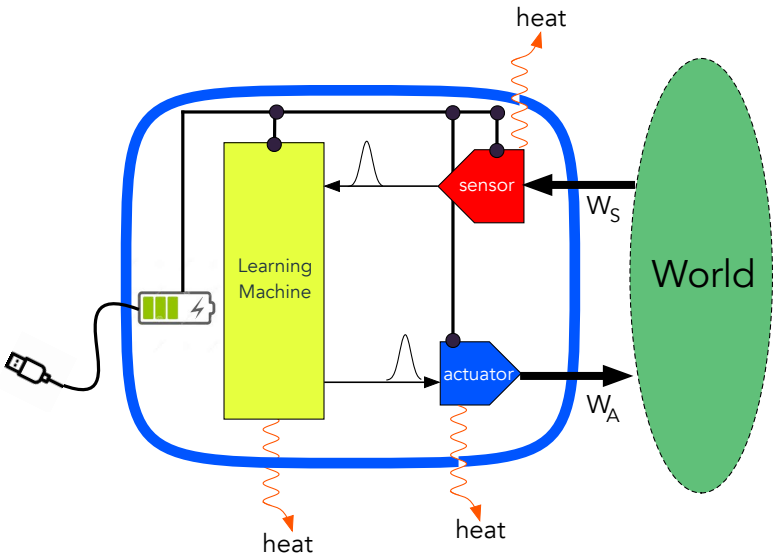

Figure 1: A minimal model of a causal agent.

This model of a causal agent functions as follows. All agents require an external source of energy, a 'battery', that does work on the agent in order to increase the agent's own *free energy*.[3] The battery can be employed to do work on the agent's specialised subsystems—maintaining them in a non-thermal-equilibrium steady state—which then dissipate heat into the environment resulting in a net increase of entropy in the environment. An actuator is a device controlled by

[2] For ease of exposition we treat these as distinct machines but in practice they may be a single machine.

[3] As a result of this, agents can only arise in a universe that is not in thermal equilibrium, and where the agents themselves are not in thermal equilibrium with their environment.

the agent that is capable of performing work on the external world, W_A, thereby decreasing the actuator's free energy, and a sensor is a device that interacts with the external world such that the world does work on it, W_S, thereby increasing the sensor's free energy. In Fig. 1 we distinguish W_A and W_S, which are intermittent, from the work done by the battery required to maintain the agent in a non-thermal-equilibrium steady state.[4]

During some interaction between the agent and the world, the world is able to do work on the agent's sensors, and they are temporarily forced away from their quiescent steady state. Similarly, the agent's actuators can be pushed away from their quiescent steady state by internal actions in the agent such that they are subsequently able to do work on the external world. These temporary excitations in the agent's subsystems—externally driven in the case of the sensors, and internally driven in the case of the actuators—can be recorded in an internal memory, or may even simply reflect an ordered series of impulses input to, or output from, the learning machine. The learning machine passes information to the actuators, and the sensors pass information to the learning machine, whereafter both sensors and actuators are restored to their quiescent steady states ready for the next round of interactions.

The role of the learning machine is to model the correlations between the actuators and sensors, and thereby learn through the agent's interaction with its environment plausible causal relations that it can attribute to external phenomena. A possible approach for the learning machine is to identify sensor and actuator actions with their corresponding states, and then the process of learning amounts to establishing correlations between these records of state. In so far as these correlations between actuator and sensor states tell the agent something about the external world, the physical state of the learning machine then embodies the causal relations between inputs to the world from the actuators and outputs from the world to the sensors. Let us consider the learning machine in more detail.

Following Milburn and Shrapnel (2020, p.16), we can envisage the learning machine of this causal agent as an *emulator* of the functional relation between the actuator outputs and sensor inputs. The learning machine compares predictions of the emulator to the actual sensor and actuator records. This is reminiscent of the general approach of Clark (2013) to understanding intelligence in animals.

Underpinning this account is the key assumption of the Church-Turing-Deutsch (CTD) principle, that every physical process can be simulated by a universal computing device (Nielsen, 2004). By assuming the CTD principle, this permits the possibility that a sufficiently complex learning machine in a causal agent can emulate *any* physical system that it happens to encounter. As a result, any such causal agent is able to model the external world as a Turing machine, where the machine provides a functional relation between its inputs, which correspond to the agent's actuator output, and its outputs, which corresponds to the sensor data that the agent receives. Learning for the agent amounts to modelling these functional relations, and by such models the causal agent *learns* the causal relations between its actions and their effects.

A simple schematic of this approach to the learning process is depicted in Fig. 2 (adapted from (Milburn and Shrapnel, 2020, p.16)). Upon the agent acting on the external world via its actuator, a primary actuator record registers the action taken. A copy of this record is sent to the emulation engine, which then generates an emulated output. We can think of this emulated output as the agent's best estimate for the data it will receive from its sensor; call this the emulation sensor record. Upon the agent receiving data from the external world via its sensor, a primary sensor record registers the sensation received. A comparator (C) then compares the primary sensor record to the emulation sensor record, with the result fed back to

[4]While we remain as neutral as possible regarding the philosophical characterisation of thermodynamic concepts—work, heat, and free energy—we take the characterisations found in Myrvold (2020) to fit naturally with our discussion.

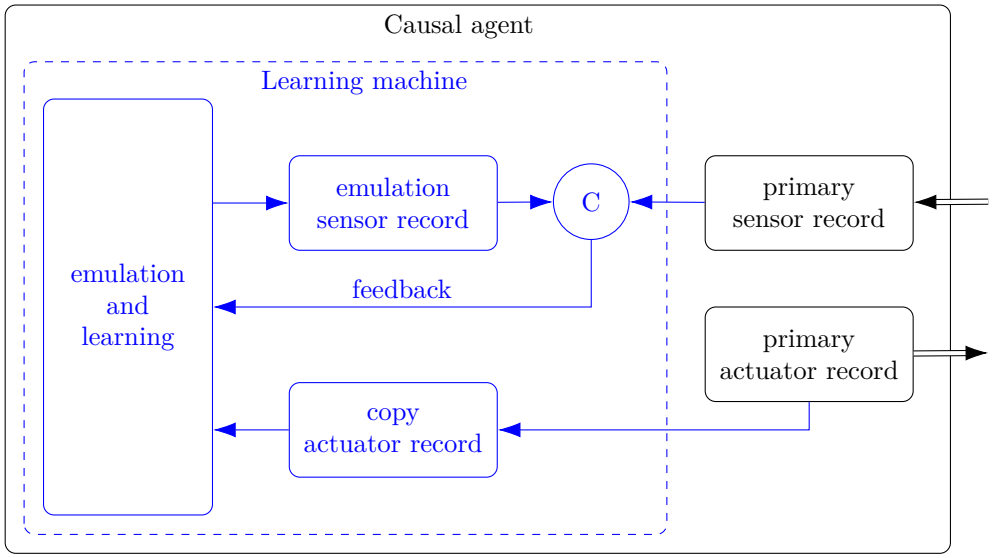

Figure 2: A schematic of a learning process based on a physical emulator with feedback. (Adapted from Milburn and Shrapnel (2020, p.16).)

the emulation engine, which then updates the engine to better emulate the sensor data. Every new action refines the emulation engine until some goal is met for the comparator output. Once the comparator goal is met, we can say that the agent has successfully developed an internal model of the functional relations governing the evolution of the actuator data to the sensor data in the external world, and so has learned the causal relations—implicit in the primary actuator and sensor records—between the variables comprising that data.[5]

Now that we have a clearer picture of this model of a causal agent, we can begin to get clearer on the role that the agent's deliberative orientation and physical constitution play in generating the causal 'perspective' of the agent. We begin this task by considering the relation between the deliberative stance of the agent and the irreversible dynamics of its network of sensors, actuators, and learning machine in the next section.

Before we do, though, it will be instructive for our purposes below to note that the actuator and sensor data, and the learning machine comprised by the emulation engine and comparator, are idiosyncratic to a particular agent. If a set of causal agents whose actuators, sensors, and learning machine operate in an identical manner are given some external physical system with

[5]This construction of a causal agent has considerable overlap with other constructions. The model for learning is similar to the concept of predictive processing developed in the philosophy of neuroscience (Kirchhoff, 2018). It echoes the emerging consensus on intrinsic networks in neuroscience (Barrett, 2017). The importance of sensors and actuators for artificial agents is a staple of textbooks on artificial intelligence (Russell and Norvig, 2010). Briegel and De Las Cuevas (2012) also stress the importance of sensors and actuators for embodied agents. Their novel concept of 'projective simulations' can be seen as a more sophisticated proposal for the learning machine in the above model. They emphasise the role of stochasticity for creative learning agents and possible quantum enhancements. Likewise, in elucidating his concept of action-based semantics, Floridi (2011) describes a two-machine artificial agent. This enables the relationship between the two-machine internal states of the artificial agent to play the role of 'semantic-inducing resources' for what would otherwise be raw bit strings without resorting to an external semantic crutch. The 'two machines' of Floridi's scheme roughly correspond to the actuator/sensor machines and the learning machine in the above model. The model presented here has significant similarities to the agent model introduced in a biological setting by Friston and Stephan (2007) and subsequent developments in Bruineberg *et al.* (2018). A robotic implementation of Fig. 2 is described in (Young, 2017).

which they can interact, one should expect that the model that each agent develops of the external physical system should consist of the same causal relations between the relevant variables with which the actuators and sensors operate. And, provided that the learning machine is effective, this similarity between models should be independent of the particular primary actuator and sensor records registered in each agent. As Milburn and Shrapnel (2020, p.16) point out: "These internal representations may well be opaque to an outside observer." However, regardless of this opacity, a set of causal agents that share actuator, sensor, and learning capabilities should share similar internal representations.[6]

We can talk, then, of an equivalence class of agents who share actuator, sensor, and learning capabilities, and thus also share internal representations of the causal relations used to model the external world. To foreshadow the argument to come in §4, we can define this equivalence class, as we did at the end of §2, as a 'perspective'; in this case a causal perspective. We emphasise that this model makes it obvious how the perspective is a function of the capabilities of the agent.[7]

3.2 The necessarily irreversible dynamics of sensors, actuators, and learning machines

Significantly for the argument we make in this work, the above model treats an agent as an open irreversible system. This means that we can best understand the key physical elements of the minimal model—that is, sensors, actuators, and the learning machine—in terms of thermodynamics. Sensors and actuators are easily distinguished by their thermodynamic properties: sensors are machines that do work on the agent's environment while the environment does work on actuators. These quantities of work are constrained by changes in the free energy of sensors (it increases) and actuators (it decreases). A technical discussion of this distinction in terms of a specific model is given by Milburn and Shrapnel (2020). For our purposes here the essential feature is that both sensors and actuators are irreversible physical systems. Neither is in thermal equilibrium but prior to sensations/actions happening they are maintained in distinct non-equilibrium steady states—quiescent states, ready to register a change.

It will be instructive for us to employ here a powerful theoretical tool from statistical physics, known as the *fluctuation-dissipation theorem* (Hoffman, 1962; Gardiner, 1983), to help in understanding the thermodynamics of a general dissipative system. The theorem explains how systems subject to dissipation of energy—for instance, due to friction—come to thermal equilibrium, whereby the average kinetic energy of a system determines its temperature. As an example of how the theorem works, consider a large molecule injected into a viscous fluid at temperature T. The molecule will experience viscous damping of its velocity, and so on average the velocity will approach zero. While the average velocity will be zero when the system is in equilibrium, the average kinetic energy, proportional to the square of the average velocity, cannot be zero as the large molecule must come to an equilibrium temperature that is the same as the temperature of the surrounding fluid bath. The fluctuation-dissipation theorem shows that the same microscopic forces responsible for viscous damping also lead to fluctuating forces on the molecule; these fluctuating forces cause the velocity of the molecule to continue to fluctuate even as the *average* fluctuation goes to zero. These fluctuations are manifest as the temperature of the molecule in equilibrium with the fluid. All dissipative systems, classical or quantum, necessarily experience

[6] We use 'similar' here since such agents will represent a causal relation via the same function up to some error defined by the stochasticity of the actuators, sensors and learning machine.

[7] We note at this point that Milburn and Shrapnel (2020) refer to two distinct descriptions of the irreversible physical systems of the agent: the "inside view", which leverages the internal physical states of the actuators and sensors from the agent's own point of view; and the "outside view", which coarse-grains the first-person description and treats the agent simply as an open physical system interacting with its environment. See (Milburn and Shrapnel, 2020) for further detail on this point.

noise in the form of thermal fluctuations of this sort. A key difference in the quantum case is that these fluctuations will occur even as the thermal noise is reduced to zero—for instance, through quantum tunnelling or spontaneous emission.

3.2.1 Sensors and actuators

As we saw above, in order to function sensors and actuators must be maintained in a given quiescent state; that is to say, with probability approaching one they are in a particular 'ready' state. The quiescent states of sensors and actuators are non-thermal equilibrium steady states with low entropy, and to achieve this they must have work done upon them by internal mechanisms within the causal agent, and so dissipate energy into the environment (as shown in Fig. 1).[8] The quiescent state cannot be maintained with certainty as that would require access to a zero entropy source. We need only arrange for sensors and actuators to be found in the quiescent state within some appropriate error bound.

By their operation, sensors and actuators are routinely and momentarily pushed away from their non-equilibrium quiescent states. From time to time, as a result of, say, a sensor's interaction with the environment, the state of the sensor will change to a new physical state with *increased* free energy, indicating that the environment has done work on the sensor. Similarly, as a result of a physical process internal to the agent, the state of an actuator will change to a new physical state with *reduced* free energy, thus doing work on the environment. In both cases energy is dissipated into the environment and the state transitions are irreversible.

3.2.2 Learning

A learning machine is necessarily an irreversible dissipative device, and this is the case regardless of the specific algorithm that it is running. It is worth first considering one basic structural limitation of learning machines. A physical learning machine can never be perfect; it must make mistakes from time to time with some error probability, ideally much less than one. To see this more clearly, consider how the entropy of a learning machine changes as it learns. Given random input into a learning machine, the initial output entropy of the learning machine is very large. As it learns, its output entropy reduces considerably. But the entropy cannot be reduced to zero—that is, the learning machine cannot master its particular task with perfection—as this would violate the third law of thermodynamics, and so a small but finite error in learning must remain.[9] Furthermore, since in a successful learning process the entropy will have dramatically reduced, the feedback training of the machine must produce a lot of heat and so a large increase in environmental entropy. As a result, learning machines are necessarily irreversible, error prone, and produce large amounts of heat and entropy on large data sets.[10] The advantage of training a learning machine, though, is that, once the machine is trained, it then has the ability to classify arbitrary inputs more efficiently.

We can explore these features of a learning machine by considering a possible physical implementation of the emulator in Fig. 2 so as to more carefully identify the dissipative elements and feedback in the system. One possible kind of emulator machine is based on a specific class of learning algorithms known as *neural network algorithms*. At the most elementary level, neural network algorithms make use of a threshold or activation function to delineate an input threshold beyond which individual neurons produce an output if the input is of sufficient intensity.

[8] In the classical case this dissipated energy is heat. In the quantum case it is simply random energy loss such as spontaneous emission into a zero temperature bath.

[9] One formulation of the third law of thermodynamics is that no finite process can cool a system down to its absolute-zero value for entropy (Nernst, 1912, p.134).

[10] A more detailed treatment along the lines summarised here is given in (Goldt and Seifert, 2017).

When implemented physically, this is a device with a highly nonlinear response; the output is a nonlinear stochastic function of the inputs. We will refer to the physical device as an 'activation switch'. It is here that dissipation, and necessarily noise, enters in a fundamental way (Goldt and Seifert, 2017, p.7).

The physical system that implements a neural network algorithm is a network of such elementary devices with the output from one forming the input to many. Clearly such a system has a strongly dissipative nonlinear dynamics and, by the fluctuation-dissipation theorem (Gardiner, 1983), must necessarily be intrinsically noisy. In a single trial, the output is a random variable with the size of the dissipation determining the scale of the fluctuations: the greater the dissipation the smaller the noise for a fixed temperature. This means that in some cases the output will not switch when it should, corresponding to an error. The objective of learning is to minimise this error by changing the weights as training proceeds and thereby changing the input to the activation switch. There is thus a trade-off between switching rate (and consequently the learning rate) and error. In order to learn, the inputs to every activation switch in the network must be varied according to a feedback rule based on how the output of an activation function varies as the input is varied. During training, work is done on the switch by the feedback control of the weights and that work is dissipated as heat.

In summary, physical learning machines, like sensors and actuators, are open dissipative nonlinear systems. They are necessarily noisy and dissipate heat as they learn. But unlike sensors and actuators, for learning machines noise is essential; the smaller the noise the slower the learning.[11] But it does not follow that physical learning machines must fail at very low temperatures as eventually quantum noise sources such as tunnelling and spontaneous emission will dominate, enabling learning to proceed even at temperatures close to zero.

4 Enriching the agent perspective

In §2 we saw two distinct aspects to causal perspectivalism. The first aspect, elucidated by Price (2007), is that the deliberative orientation of agents such as ourselves is a function of the significantly constrained epistemic access that agents have to their environment. For thermodynamically biased, inherently directed agents such as us, where the events about which we deliberate are unknown and open to manipulation, our causal attributions are inviolably directed from the known to the unknown parts of our environment. This 'directedness' constitutes our causal perspective by underpinning the distinctions we make between exogenous and endogenous variables, and so constraining the direction of exploitation of the functional relations between variables.

The second aspect, delineated by Evans (2020) (see also Baron and Evans (2021)), is that the causal relations that an agent attributes to the external world are relative to a system-environment split and a coarse-graining, both of which are heavily constrained by the particular physical capabilities of the agent for intervening on and detecting the world. This agent-centric process of manipulating and modelling the world comprises a further epistemic constraint on the access agents such as ourselves have to our environment—agents are limited by the constitution of their actuators and sensors, in the terminology of §3. As such, the physical capabilities of an agent contribute to the agent's causal perspective.

These natural limitations on a causal agent may very well constrain the epistemic access that an agent can gain about its environment, but they also equally define that epistemic access, and in so doing enrich the agent perspective with content from the external world. Moreover, any such constraints will be idiosyncratic to a kind of agent—one that is temporally embedded,

[11] For a more precise formulation of this principle see Eq. 17 of (Goldt and Seifert, 2017).

or with particular physical capabilities—and so define an equivalence class of agents that share a perspective. An agent, or set of agents, with a vastly different network of actuator, sensor, and learning machine might arrive at a completely different model after interacting with the very same physical system. Let us consider how the above model of a causal agent illustrates causal perspectivalism before going on to demonstrate how objectivity can arise from a shared perspective.

4.1 An illustration of causal perspectivalism

Regarding the manner in which an agent's epistemic access is constrained by the deliberative orientation of the agent, Milburn and Shrapnel (2020) argue explicitly that their model provides an illustration of how the asymmetry of intervention and the asymmetry of learning lead to a constrained agential perspective from which asymmetric causal attributions are made. The arguments in this section aim to flesh out this claim in greater depth.

The 'state of knowledge' of the causal agent modelled in §3 consists of a model of the functional relations between the actuator and sensor data. The model is developed by the agent's learning machine, and we take it to represent the agent's own causal model of the external physical system with which it is interacting. It is straightforward to see that this epistemic state is heavily constrained by the agent's own inherently directed deliberative stance. Firstly, as noted above, the distinction between actuators and sensors is a thermodynamic one: actuators do work on the environment while the environment does work on sensors. As a result, actuators and sensors are necessarily irreversible dynamic systems that require a low entropy source of energy to operate, driven by internal mechanisms within the causal agent.

Secondly, the dynamics of learning is also irreversible and necessarily noisy. Noise plays a fundamental role in physical learning machines: in classical physics, if we were to reduce the noise, which can only be achieved by reducing the temperature of the environment, we would fundamentally reduce the rate of learning of the agent. Thus, learning inevitably generates heat and entropy in the agent's environment and the more heat and entropy it can generate the better it can learn. As a result of this, the process of learning itself is constrained by the same factors that constrain the agent's own inherently directed deliberative orientation.

We take it that the causal agent's epistemic access to dynamical behaviour in the physical external world is gained through the process of interventions, detection, and learning. Due to the thermodynamic directedness of actuators, sensors, and learning, any such epistemic access to the world must be thermodynamically downstream from the actuator and thermodynamically upstream from the sensor. In simple terms, the agent can only act towards the future, and can only gain knowledge of the past. This is surely what Price means when he says, "the notion of intervention is ineliminably perspectival" (Price, 2007, p.269). Thus, we can see clearly that the deliberative orientation of the agent is necessarily aligned with the thermodynamic directedness of its actuators, sensors, and learning machine. This heavy constraint inescapably contaminates the nature and utility of the agent's causal model—the agent's 'knowledge'—to attribute causal relations to the external world. The model only has utility as a guide to action that exploits modelled—or 'known'—functional relations, and those relations are directed from variables manipulable by the actuator (exogenous), and detectable by the sensor (endogenous). We ordinarily take these relations to be temporally directed, which, if our thinking is right here, is a clear function of the deliberative stance of the agent, and so the thermodynamic bias of the agent's specialised subsystems. Thus the directedness of the network of actuator, sensor, and learning machine is embodied in the directedness of the causal model.

The thermodynamically biased epistemic access that the causal agent has to the external physical world constitutes what we have labelled above a perspective: the agent is situated

and embedded in an environment that is inescapably thermodynamically oriented, due to the deliberative orientation of the agent's own network of actuator, sensor, and learning machine. We take this deliberative orientation to be the temporal direction. This model of a causal agent then provides an illustration of the first aspect of causal perspectivalism, that our causal attributions are inviolably temporally directed as a function of our thermodynamically biased perspective.

Next, let us consider the manner in which the learned model of the causal agent is relative to a system-environment split and a coarse-graining. There is a sense in which, given a sufficiently complex range of actuator and sensor capabilities, an agent has a specific modelling *choice* to make about the distinction between system and environment and the level of grain at which appropriate dynamical variables can be objectified. But, in another sense, an agent *cannot* choose the range of actuator and sensor capabilities. And so choices concerning a system-environment split and a coarse-graining outside of this range are simply not possible. In this sense, then, the agent is constrained to interact with the world, and make modelling choices, only according to the set of dynamical variables that are manipulable-and-detectable by the actuator and sensor system (in addition to being functionally related in the actuator and sensor data).[12]

Considering again that the state of knowledge of our causal agent from §3 is a causal model of the functional relations between the actuator and sensor data, we can see that this causal model will only contain functional relations that are exploitable according to the physical constitution of the actuator and sensor. For instance, imagine an actuator and sensor pair that consisted of a photon source and photo sensitive detector. The exploitable functional relations for the agent will consist of only the photonic structure in the world (which, in the classical case, would be electromagnetic phenomena). The agent is 'blind' to any other structure. Thus the causal model—the agent's epistemic state—is constrained in precisely this way: the agent will develop a model of the photonic structure of the world, attributing causal relations that are exploitable for the agent between manipulable-and-detectable variables, the set of which is defined by the physical constitution of the sensor and actuator, but be blind to any other external structure outside of this model.

Just as we considered above with the case of the deliberative orientation of the agent, the physical capabilities of the agent generate an inescapable environment within which it is situated and embedded. This environment is not exactly given by the external physical world, but is a function of the physical constitution of the agent's own network of actuator, sensor, and learning machine.[13] This situatedness once again defines a perspective, and so this model of a causal agent provides an illustration of the second aspect of causal perspectivalism, that our causal attributions are inviolably constrained by our own physical constitution.

This analysis of the perspectivalism generated by the physical constitution of a causal agent suggests a further interesting consequence. We can think of a scientific instrument that an agent employs to explore the external world as an extension of the agent's actuators and sensors. Such instruments extend the range of possible interventions beyond the capability of the actuators and increase the types of phenomena that can be detected by the sensors. Thus, just as the physical constitution of an agent's actuators and sensors define the perspective of the agent, so too does the physical constitution of a scientific instrument extend the scope of that perspective. However, it should be noted that the use of scientific instruments cannot enable an agent to break free of its perspective, only to redefine the scope of the perspective. (There is, of course, on this view no possibility of a perspective-free interaction with the external world.) Thus using a scientific instrument to intervene on the world requires the instrument to be such that we

[12]This claim is also made by Giere (2006, p.14): "instruments are sensitive only to a particular kind of input. They are, so to speak, blind to everything else."

[13]Employing the terminology of Kirchhoff (2018), we can imagine that the agent's own 'Markov blanket' is an environment of sorts with which its internal states must contend.

can intervene accordingly on it, and using an instrument to detect some property of the world requires the output of the instrument to be detectable by us.[14]

4.2 Intersubjective objectivity

As we noted in §3.2.2, the specific network of actuator, sensor, and learning machine that constitutes a particular causal agent is idiosyncratic to that agent. A set of causal agents who share identical networks of actuator, sensor, and learning machine, and who explore some external physical system should converge on the same learned models of that system (within some error bound dictated by the comparator). By sharing such physical capabilities, this set of agents share a perspective. A perspective thus defines an equivalence class of agents who will all agree (within some bound) on their representation of the external physical world.

With this in mind, we can now connect this discussion to the one at the end of §2 concerning the objectivity of causal relations. Recall Price's claim that objectivity is dependent upon a particular perspective. And since our own human perspective is a function of inescapable epistemic constraints, the perspective is stable across human agents. This then underpins strong intersubjective agreement between agents sharing a perspective—we all generally agree on our attribution of causal relations in the world—and so leads to a sense of 'objectivity', where all agents necessarily agree on their best-practice representations of the physical world. Thus we can see now that the above model of a causal agent provides an illustration of the intersubjective objectivity that is a key part of causal perspectivalism. The model allows us to be even more specific about this intersubjectivity: it occurs only within a shared perspective. That is, it is simultaneously intersubjective and 'intra-perspective' objectivity

It is interesting to note some features of the model of a causal agent, and how that translates into this notion of objectivity in causal perspectivalism. The functional relations that characterise the causal relations identified by some agent in the actuator and sensor records describe exclusively the steady states of the learning machine inside the agent. They do not in fact describe anything in the world, although we can imagine a causal agent making causal claims about the world based on its internal representation. Indeed, it may well be that the world cannot be described by recursive functions of the kind that learning agents are constrained to. (For instance, it may well be that the CDT principle is false.)

This feature of the model of a causal agent dovetails nicely with causal perspectivalism in the sense that causal perspectivalism is a kind of *causal republicanism*, which Price (2007) describes as an "irenic third way" between causal realism and causal eliminativism (Russell, 1913). According to causal republicanism, the utility, or even indispensability, of the notion of causation for modelling the world does not undermine the claim that causal concepts are constructed by us (Price and Corry, 2007). Causal claims, on this view, "essentially involve a projection onto the world of features of our perspective as deliberative agents" (Price and Corry, 2007, p.4). Thus we can resurrect Woodward's complaint from above that Menzies and Price "are not very forthcoming about just what is meant by their claims that causation is "projected" onto the world by us or "constituted" by our beliefs and attitudes". We can see now by way of the above illustration precisely what is meant by this 'projection': the modelled functional relations between the actuator and sensor records are projected as causal relations onto the relevant external physical systems, and these causal relations can be considered perfectly objective amongst agents who share a perspective (as all human agents do).

[14] In unison with these claims, Giere (2006, p.126) states that "an observer, whether a human or an instrument, can interact with an object only from the observer's own particular perspective". Evans (2020, p.16) expands on this sentiment by pointing out that our instruments are necessarily designed and engineered to concatenate with our own physical capabilities. See also (Creţu, 2020).

5 The final picture and future directions

We framed the argument of this work as a counterpoint to the criticism that agency accounts are unacceptably subjective or anthropocentric. We have demonstrated how a suitably competent nonhuman agent could generate robust causal knowledge of the world, thereby reducing the significance of specifically 'human' agency in agency accounts, and consequently how we could come to think that causal relations would be different if humans had been different. The model of a causal agent considered here provides a clear illustration of how causal relations can be a function of the agent perspective, and how such causal relations can be 'objective' within a shared perspective, defining an equivalence class of agents. Thus, far from agency accounts being unacceptably subjective or anthropocentric, we claim that the model shows that agency accounts are perfectly capable of characterising causation in the absence of human agents (though crucially not in the absence of *any* agent), and can achieve a kind of objectivity as a function of intersubjective agreement on causal ascriptions.

As we argued above, the deliberative orientation of the agent is necessarily aligned with the thermodynamic directedness of its actuators, sensors, and learning machine. As a result, the agent's causal ascriptions to the world only have utility as a guide to action that exploits functional relations directed from exogenous to endogenous variables. We noted above that we ordinarily take these relations to be temporally directed. This account provides a neat narrative for why three familiar asymmetries are aligned with the causal asymmetry: the temporal direction from exogenous to endogenous variables is a function of the deliberative stance of the agent, and thus the thermodynamic bias of the agent's specialised subsystems.

Moreover we briefly considered the sense in which the agent's internal conception of time arises within the agent's network of subsystems. We made no assumption that there must be a global temporal direction with which the agent's internal time aligns. Recall, however, that agents can only arise in a world that is not in thermal equilibrium, so there must be a local thermodynamic gradient providing an external low entropy source of energy to the agent. But this thermodynamic arrow need not be global. Different pockets in the universe may harbour irreconcilable thermodynamic arrows, and so harbour agents who would not agree on the direction of causal relations or the arrow of time. But it is clear from our arguments above that such agents would trivially occupy different causal perspectives, illustrating a feature rather than a defect of causal perspectivalism. We can then state that agents inhabiting a pocket of the universe with a single consistent thermodynamic gradient will share a causal perspective, and will similarly share a direction to their internal conception of time.

As a final note, it is clear that the model of a causal agent presented here is a simplification. We would like to finish by providing some brief speculation concerning two ways that this simplified model of a causal agent might be extended. In the first instance, it would be useful to expand the capabilities of the agent—for example, by adding arrays of sensors and actuators that make use of distinct physical effects, along with associated learning machines. A very important class of actuators would simply move the agent around in physical space changing its environment, and thus the response of its sensors. Furthermore, one could even imagine adding another *layer* of internal sensors, actuators, and learning machines responding directly to the devices one layer down, and making choices between them so as to optimise the thermodynamic efficiency at the lower level (Young, 2017). Models like this could learn causal relationships between causal relationships. This nesting could continue to higher levels, leading to a self-referential structure enabling an agent to learn and emulate itself (modulo any Gödellian argument).[15]

Another extension to this simplified model would be to consider a *quantum* causal agent.

[15]This idea has similarities to—albeit fringe in the philosophy of mind—self-representational approaches to consciousness (Hofstadter, 1999; Kriegel and Williford, 2006; Kriegel, 2009).

So far the discussion of this work has been generally directed towards classical agents; that is, we have considered agents whose network of actuators, sensors, and learning machine consist of classical devices. It is an interesting open question how this framework for a causal agent could extend to the case of a quantum agent. There may well be no accepted precise definition of a quantum agent, but we surmise that such an agent would be defined by the fact that its actuators, sensors, and learning machine were quantum devices, and the internal processing of the data it gathers by these devices may well proceed as a coherent quantum computation. We make here two passing comments on such a quantum causal agent.

Firstly, we noted above that causal agents cannot operate *in principle* without error. This is a direct consequence of the fact that learning machines become inefficient as noise is eliminated. All dissipative systems are necessarily subject to noise in the form of random perturbations to their dynamics. In the classical case this noise is thermal in origin. However, in the quantum case it arises from pure quantum events, such as tunnelling and spontaneous emission occurring, even at zero temperature.[16] So whereas at zero temperature the rate of learning of a classical casual agent goes to zero, a quantum causal agent will continue to learn. This may provide scope for tremendously more energy efficient learning algorithms in the future.

Secondly, while grounding the objectivity of causal relations in the intersubjective agreement of agents within a perspective is, for classical causal agents, perfectly consistent with a perspective-independent external reality, this is not so for quantum causal agents. For quantum causal agents, the agent perspective—that is, the role of the 'observer'—plays an ineliminable role in objectifying quantum systems (Evans, 2020). For example, an agent using electric field sensors and actuators for probing and detecting light phenomena will learn a different set of causal relations to an agent using photon counting, even when embedded in the very same environment. (This may even shed new light on old arguments about wave-particle duality.) Thus, it seems doubtful that the causal model of a quantum causal agent can be made consistent with a perspective-independent (observer-independent) quantum reality. Nonetheless an intersubjective objectivity would appear to be possible even in this case given the discovery of quantum theory.

Acknowledgments: For useful discussion and comments we would like to thank Sam Baron, and the audience at the Quantum Causal Agents workshop hosted by the University of Queensland in 2021. We acknowledge support from the University of Queensland. GJM and SS wish to acknowledge the support of the Foundational Questions Institute (FQXi-RFP-1814). PWE acknowledges support from the Australian Government through the Australian Research Council (DE170100808).

References

Anderson, E. (2019). Feminist Epistemology and Philosophy of Science. *In* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University.

Baron, S. and Evans, P. W. (2021). What's So Spatial about Time Anyway? *The British Journal for the Philosophy of Science* **72**(1): 159–183. doi:10.1093/bjps/axy077.

Barrett, L. F. (2017). *How Emotions Are Made: The Secret Life of the Brain.* Houghton Mifflin Harcourt Books.

Beebee, H. (2007). Hume on Causation: The Projectivist Interpretation. *In* H. Price and R. Corry

[16] Both quantum tunnelling and spontaneous emission arise as a direct result of quantum entanglement between a subsystem and its environment.

(eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford, chapter 9, pp. 224–249.

Briegel, H. J. and De Las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports* **2**: 400. doi:10.1038/srep00400.

Bruineberg, J., Kiverstein, J. and Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese* **195**: 2417–2444. doi:10.1007/s11229-016-1239-1.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**: 181–204. doi:10.1017/S0140525X12000477.

Creţu, A.-M. (2020). Perspectival Instruments. philsci-archive.pitt.edu/18341/.

Evans, P. W. (2020). Perspectival objectivity. *European Journal for Philosophy of Science* **10**(2): 19. doi:10.1007/s13194-020-00286-w.

Floridi, L. (2011). *The Philosophy of Informationn*. Oxford University Press, Oxford.

Friston, K. J. and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* **159**: 417–458. doi:10.1007/s11229-007-9237-y.

Gardiner, C. W. (1983). *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer, Berlin, Heidelberg.

Giere, R. N. (2006). *Scientific Perspectivism*. University of Chicago Press, Chicago. doi:10.7208/chicago/9780226292144.001.0001.

Goldt, S. and Seifert, U. (2017). Thermodynamic efficiency of learning a rule in neural networks. *New Journal Physics* **19**: 113001. doi:10.1088/1367-2630/aa89ff.

Hoffman, J. G. (1962). The fluctuation dissipation theorem. *Physics Today* **15**: 30. doi:10.1063/1.3057971.

Hofstadter, D. R. (1999). *Godel, Escher, Bach: An Eternal Golden Braid*. Basic Books, Inc., New York.

Hume, D. (1748). *An Enquiry Concerning Human Understanding*. A. Millar, Strand, London.

Ismael, J. (2016). How do causes depend on us? The many faces of perspectivalism. *Synthese* **193**: 245–67. doi:10.1007/s11229-015-0757-6.

Kirchhoff, M. D. (2018). Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philosophical Studies* **175**(3): 751–767. doi:10.1007/s11098-017-0891-8.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford University Press, Oxford.

Kriegel, U. and Williford, K. (eds.) (2006). *Self-representational Approaches to Consciousness*. MIT Press.

Menzies, P. and Price, H. (1993). Causation as a Secondary Quality. *The British Journal for the Philosophy of Science* **44**(2): 187–203. www.jstor.org/stable/687643.

Milburn, G. J. and Shrapnel, S. (2020). Physical grounds for causal perspectivalism. arXiv:2009.04121 [physics.hist-ph].

Myrvold, W. C. (2020). The Science of $\Theta\Delta^{\mathbf{cs}}$. *Foundations of Physics* **50**: 1219–1251. doi:10.1007/s10701-020-00371-3.

Nernst, W. (1912). Thermodynamik und spezifische Wärme. *In Sitzung der physikalischer-mathematischen Classe am 1. Februar*. Königlichen Akademie der Wissenschaften, Berlin, Sitzungsberichte Königlich Prüßischen Akademie der Wissenschaften, pp. 207–216.

Nielsen, M. (2004). Interesting problems: The Church-Turing-Deutsch Principle. michaelnielsen.org/blog/interesting-problems-the-church-turing-deutsch-principle.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition.

Pfeifer, R. and Bongard, J. (2006). *How the Body Shapes the Way We Think : A New View of Intelligence*. MIT Press, Cambridge, Massachusetts.

Pfeifer, R. and Gabriel, G. (2009). Morphological Computation – Connecting Brain, Body, and Environment. *In* B. Sendhoff, E. Körner, H. Ritter, K. Doya and O. Sporns (eds.), *Creating Brain-Like Intelligence: From Basic Principles to Complex Intelligent Systems*, Springer, Berlin Heidelberg.

Price, H. (2007). Causal Perspectivalism. *In* H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford, chapter 10, pp. 250–292.

——— (2017). Causation, Intervention and Agency: Woodward on Menzies and Price. *In* H. Beebee, C. Hitchcock and H. Price (eds.), *Making a Difference: Essays on the Philosophy of Causation*, Oxford University Press, Oxford, chapter 5, pp. 73–98.

Price, H. and Corry, R. (2007). A Case for Causal Republicanism? *In* H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford, chapter 1, pp. 1–10.

Russell, B. (1913). On the Notion of Cause. *Proceedings of the Aristotelian Society* **13**: 1–26. www.jstor.org/stable/4543833.

Russell, S. J. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, New Jersey.

Slezak, M. (2009). *The Concept of Cause: A Case for Subjectivism*. Ph.D. thesis, University of Sydney.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

——— (2007). Causation with a Human Face. *In* H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press, Oxford, chapter 4, pp. 66–105.

——— (2009). Agency and Interventionist Theories. *In* H. Beebee, C. Hitchcock and P. Menzies (eds.), *The Oxford Handbook of Causation*, Oxford University Press, Oxford, chapter 11, pp. 234–262. doi:10.1093/oxfordhb/9780199279739.003.0012.

Young, R. (2017). A General Architecture for Robotics Systems: A Perception-Based Approach to Artificial Life. *Artificial Life* **23**: 236.