European Journal of Neuroscience

WILEY

EJN European Journal of Neuroscience FENS

SPECIAL ISSUE ARTICLE OPEN ACCESS

The Relevance of a Philosophical Toolkit to Advance Neuroscience

Beyond Mechanism—Extending Our Concepts of Causation in Neuroscience

Henry D. Potter | Kevin J. Mitchell

Smurfit Institute of Genetics and Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland

Correspondence: Kevin J. Mitchell (kevin.mitchell@tcd.ie)

Received: 26 June 2024 | **Revised:** 24 February 2025 | **Accepted:** 2 March 2025

Associate Editor: Igor Branchi

Funding: HP was supported by a Provost's Fund grant (1481.9050961) from Trinity College Dublin.

Keywords: agency | causal explanation | historicity | reductionism | top-down causation

ABSTRACT

In neuroscience, the search for the causes of behaviour is often just taken to be the search for neural mechanisms. This view typically involves three forms of causal reduction: first, from the ontological level of cognitive processes to that of neural mechanisms; second, from the activity of the whole brain to that of isolated parts; and third, from a consideration of temporally extended, historical processes to a focus on synchronic states. While modern neuroscience has made impressive progress in identifying synchronic neural mechanisms, providing unprecedented real-time control of behaviour, we contend that this does not amount to a full causal explanation. In particular, there is an attendant danger of eliminating the cognitive from our explanatory framework, and even eliminating the organism itself. To fully understand the causes of behaviour, we need to understand not just what happens when different neurons are activated, but *why those things happen*. In this paper, we introduce a range of well-developed, non-reductive, and temporally extended notions of causality from philosophy, which neuroscientists may be able to draw on in order to build more complete causal explanations of behaviour. These include concepts of criterial causation, triggering versus structuring causes, constraints, macroscopic causation, historicity, and semantic causation—all of which, we argue, can be used to undergird a naturalistic understanding of mental causation and agent causation. These concepts can, collectively, help bring cognition and the organism itself back into the picture, as a causal agent unto itself, while still grounding causation in respectable scientific terms.

1 | Introduction

What causes a behaviour to occur? This is a central question at the heart of several major topics in philosophy, including the problems of free will and agency. It also represents one of the main explanatory objectives of the field of neuroscience: Modern neuroscience seeks to explain behavioural phenomena by developing an understanding of how the brain generates behaviour. This objective typically relies on three basic assumptions. First, that behaviour (and cognition) are underpinned in some way by neural activity. Second, that this 'underpinning' relationship is to be understood in causal terms. And, third, that to explain a phenomenon, such as the occurrence of a particular behaviour, is to identify and cite its causes—a view known in the philosophy of science as a causal theory of explanation (Woodward 2005).

Identifying reliable causal relationships both within the brain, and between brain and behaviour, is therefore often taken to be a central project of the field of neuroscience. In a recent article on the topic, Ross and Bassett state: 'A central aim of neuroscientific research is to clarify the causal structure of the brain, be

https://doi.org/10.1111/ejn.70064

that at the lower scales of molecular and cellular interactions or the higher scales of neural circuitry, brain regions and macroscale networks' (Ross and Bassett 2024, p. 82). Similarly, Barack and colleagues state: 'In neuroscience, we are often interested in things like the events in the brain that "cause" behavior or the events in the brain that "cause" other brain events' (Barack et al. 2022, p. 654), with the motivation being that identifying the relevant neural causes will enable us to then *explain* the occurrence of the neural or behavioural event in question.

The standard approach to understanding how the brain generates behaviour is therefore one of searching for the neural mechanisms of behaviour. As Ross and Bassett describe: 'it is common to find claims that genuine explanations in neuroscience always require the elucidation of mechanistic information about the brain, where mechanistic information is understood as lower-scale causal detail that produces the brain outcome of interest' (2024, p. 82; see also Gomez-Marin 2017). Under this view, understanding the causes of a behaviour just *is* elucidating the underlying mechanism(s) whereby the activity of single neurons, neural circuits, or neural populations *causes* or *brings about* the behavioural outcome of interest.

This approach is implicit in neuroimaging research, for example, where the aim is often to identify the neural *correlates* of specific behaviours or mental states in humans or other organisms. These are then commonly (if often tacitly) taken to be the candidate *causes* of the behaviour (or mental state) in question. Likewise, in lesion studies, the aim is often to provide complementary evidence to support this program of mechanistic localisation and decomposition (Silberstein and Chemero 2013; Silberstein 2021) by showing not only that some area is active during a behaviour (such as episodic memory or face detection or speech), but that the area is *required* for that behaviour to occur.

The search for the neural causes of behaviour received a major boost with the invention of optogenetic technologies in 2005 (Boyden et al. 2005). These technologies, alongside other experimental manipulation techniques (such as pharmacology and transcranial magnetic stimulation, for example), allow researchers to directly intervene on the activity of specific neural elements (be that individual neurons, neural pathways, circuits, or whole populations), in order to test whether changes in that neural element lead reliably to changes in a given behaviour or mental state (Kim et al. 2017).

This interventionist approach is seen as the gold standard for detecting genuine causality in the world (Pearl 2009; Woodward 2005), and it has led to some striking results in the study of organismal behaviour. Using these techniques, researchers have been able to identify neural states that appear both 'necessary' and 'sufficient' for a specific behaviour, such as an avoidance behaviour, to occur (e.g. Castaneda et al. 2024; Filipowicz et al. 2022; Siemian et al. 2021; cf. Yoshihara and Yoshihara 2018; Gomez-Marin 2017). Sufficient in the sense that, when the neural state is optogenetically induced, it reliably results in the specific behaviour or change to cognitive operations, even in incongruent contexts. And necessary in the sense that, when the neural element is inhibited from firing, through inactivations or lesions, the behaviour seems to be impeded, thereby indicating that the neuron, circuit, or brain region is also *required* for the behaviour to occur (Kim et al. 2017). With this information, researchers have effectively been able to exert control over an animal's behaviour, simply by activating or inactivating the identified neural mechanism.

An additional inference is that when an animal is going about its normal business in the natural world, its behaviour is similarly *being caused by* the firing patterns of these neurons, in a way that allows us to not only successfully explain the occurrence of this type of behaviour under laboratory conditions, but also to explain its occurrence in more naturalistic settings. As articulated by Deisseroth and colleagues: 'This integrated approach now supports optogenetic identification of the *native*, necessary and sufficient causal underpinnings of physiology and behavior on acute or chronic timescales and across cellular, circuit-level or brain-wide spatial scales' (Kim et al. 2017, p. 222, our emphasis in bold).

The question for our purposes is: how should we interpret these findings? What do they tell us with regard to our original question of 'what causes a behavior to occur?' Faced with the evidence of optogenetic control over animal behaviour, it is hard to resist the rather stark impression that the manipulated neural variables are *the* explanatorily relevant causal elements of the behaviour in question—They are what is 'responsible for' or 'in control of' that effect. In particular, the capacity to exogenously *control* behaviour in real-time, by activating some neurons or other, strongly creates the impression that one has successfully identified the primary *causes* of that behaviour. After all, if we understood how the brain generates a behaviour well enough to be able to control the behaviour through neural manipulations, one might wonder what else is there left to understand? (cf. Krakauer et al. 2017).

We call this view of causation within the brain, and between brain and behaviour, the '*driving*' *view of causation*, as it is what is implied by the driving metaphor that is commonly used to describe the results of these optogenetic studies. Consider, for example, the recent discovery 'that a subpopulation of LH [lateral hypothalamus] GABAergic neurons … specifically drives appetitive behaviors in mice' (Siemian et al. 2021, p. 1). Or several recent studies that have impressively applied systematic optogenetic activation of individual neurons or sets of neurons across the nematode and fruitfly nervous systems to derive the map of responses that follow from each such activation. In the case of the nematode C. elegans, for example, this work is presented as a 'neural signal propagation atlas' (Randi et al. 2023). The authors describe how 'direct measures of signal propagation allow us to define mathematical relations that describe how the activity of an upstream neuron drives activity in a downstream neuron' (p. 406), or, more broadly, 'how a stimulus in one part of the network drives activity in another' (p. 413).

Similarly, in the case of the fruit fly brain (Pospisil et al. 2024), the authors state that: 'A long-standing goal of neuroscience is to obtain a causal model of the nervous system. This would allow neuroscientists to explain animal behavior in terms of the dynamic interactions between neurons'. (p. 2). By systematically optogenetically stimulating different regions and recording the consequences, the authors claim to be able to move beyond the static connectome and model what they call the ‘effectome’, or ‘causal model of the fly brain’: a model of the activity that each node of the network *drives* into effect, when activated.

Using a driving metaphor to conceptualise neural activity in this way reflects a legacy of foundational work on simple reflex systems, which are both the origin of our initial insights into neural signalling (Sherrington 1910) and the entry point for many introductory neuroscience texts (as discussed by Brembs 2021; Buzsáki 2019; Cisek 1999; Cobb 2020; Dewey 1896). In these systems, a sensory signal is detected and a series of neural relays is initiated, with each element *driving* the activity of the succeeding one, like dominos in a chain, until a pre-determined behaviour results.[1] In seeking to apply this conceptual framework to other systems in the brain, there is perhaps a sense that the logic of these simple reflex circuits can simply be scaled up and complexified, to explain what goes on at the level of larger neural systems, or even the whole brain.

This entirely feedforward, driving view of causation was articulated as early as 1890, by William James, who wrote that ‘The whole neural organism, it will be remembered, is, physiologically considered, but a machine for converting stimuli into reactions’ (James 1890, pp. 372; as quoted by Brembs 2021). The suggestion is that, in almost every case, the appropriate way to understand a neuron's firing activity is as following inevitably or passively from the firing of a neuron upstream of it: The activity in upstream neurons drives or necessitates the activity we see in downstream neurons. And thus, ultimately, it drives and necessitates the eventual behaviour.

For our purposes, the crucial upshot of this ‘driving’ view of causation in the brain, within the context of neuroscience's search of neural mechanisms, is that it paints a picture of the causes of behaviour that is inherently reductive in three key ways (Figure 1). First, it suggests a *vertically reductive* perspective in which, while one might conveniently and even effectively *describe* the processes of behavioural control in terms of mental states, like beliefs or desires, or cognitive operations or decisions, these are not seen as the right level for a truly *causal* explanation. Instead, it is the so-called neural ‘vehicles’ of these states (i.e. the activity of neural *mechanisms*) that are taken to be doing the ‘real’ causal work in *driving* the downstream behavioural effect. From this perspective, mental and cognitive states are *explained away* as mere epiphenomena; there is simply no room left in the causal schema for anything like the agent's conscious deliberations, or even just its cognitive processes per se, to make any sort of difference to how it behaves.

Second, this approach entails a *horizontally reductive* perspective in that it assumes that we can decompose the nervous system into various neural parts and isolate the explanatorily relevant causes of any specific behaviour to the activity of *just some of those parts*, allowing us to effectively ignore its wider neural context. From this perspective, the organism itself—as a causal agent—recedes from view or even disappears entirely from causal explanations of its own behaviour (Franklin 2014; Potter and Mitchell 2022). Even when these explanations are couched at the level of extended circuits and larger systems, the sense remains that the behaviour of the organism at any moment is simply being controlled by a subset of its neural parts.

Lastly, and less obviously, such approaches also imply a view of behaviour that is *temporally reductive*. Viewing an organism's behaviour as being *driven* into action primarily by the activation of a specific neural mechanism strongly implies that all one needs to know about the causes of a given behaviour is the currently active patterns of neural activity. From this perspective,

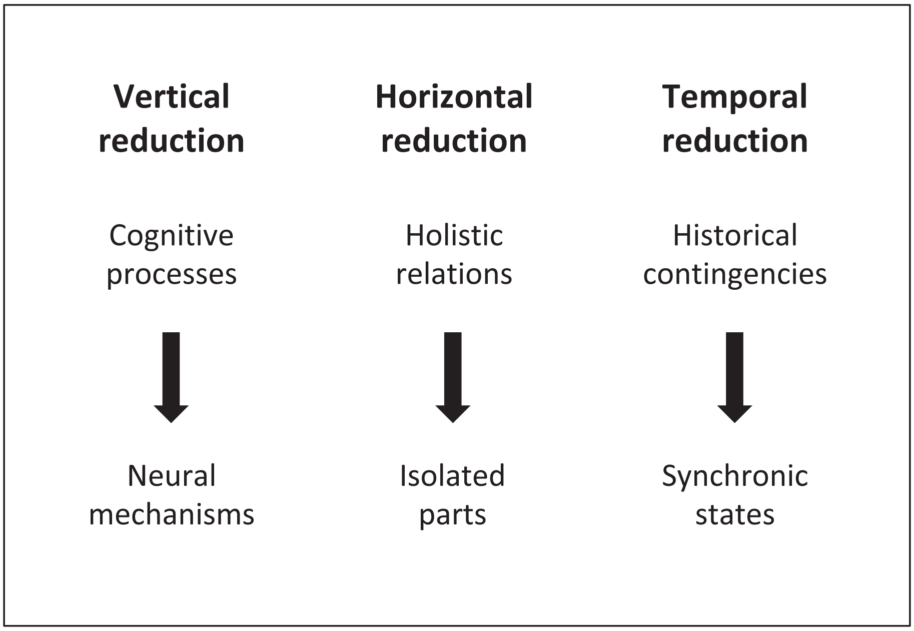

FIGURE 1 | Varieties of causal reduction. Taking neural mechanisms to be the explanatorily relevant causes of a behaviour entails a vertical reduction in ontological levels, from the cognitive to the neural, a horizontal reductionism, involving isolation and decomposition, and a temporal reduction, with an exclusive focus on synchronic states.

behaviour is depicted as the outcome of an entirely Markovian neural process. Neither the historical context that shaped these neural processes, nor the organism as a *diachronic* entity with extension and continuity in time, are considered relevant to the causal explanation of its behaviour.

In this paper, we argue that the reductive focus on synchronic neural mechanisms provides an incomplete and misleading way to think about the causes of behaviour because it relies on a needlessly narrow conception of causality. When we only have reductive, synchronic frameworks for thinking about causation, such as the driving metaphor, then it is inevitable that we will only see reductive, synchronic answers to our question of 'what causes a behavior?'. These will necessarily be ones that tend to eliminate the organism itself from the causal picture. Crucially, such a view ignores the fact that the patterns of neural activity *mean something* to the organism and that the causality in the system depends on that meaning.

Here, we introduce a range of well developed, non-reductive and temporally extended notions of causality from philosophy, which neuroscientists may be able to draw on in order to bring the organism back into the picture, as a causal agent unto itself, while still grounding causation in respectable scientific terms. In particular, we argue that a full understanding of causation in living organisms requires a diachronic view, extended through time, which centres the *meaning* of neural states. Such a view offers ways to understand the relationship between cognitive and neural processes without eliminating the former or reducing them to the latter.

2 | Production and Dependence Causes

A common folk conception of causation simply equates causes with physical forces. On this view, a cause is an event that *produces* an outcome through a transfer of energy—what List and Menzies call some causal 'oomph', as in one billiard ball hitting another (List and Menzies 2017). This is known in the philosophical literature as a 'producing' *notion of causation* (Hall 2004) and we can see echoes of this view in the 'driving' language employed in the examples above (even though synaptic transmission does not in fact involve a transfer of energy, per se, or of any physical force).

An alternative conception of causation, popular in the philosophical literature, is a broader notion known as 'difference-making' or 'dependence' causation (Hall 2004; Woodward 2005; Pearl 2009; List and Menzies 2017; Barack et al. 2022). Under this view, causes are thought of as counterfactual *difference-makers*—that is, a cause is taken to be any variable that could have changed how some event unfolded, had it been different to how it actually was. This captures the intuition that when we think of A as a cause of B happening, we usually mean that if A had not been the case, B would not have occurred.

The difference-making notion of causation clearly includes within its remit the producing (or 'driving') causes that supply some 'oomph' in bringing about an outcome, but it also makes room for a much wider range of conditions and factors to count as causal—those that *also* had to obtain in order for the producing cause to have the effect it did. Consider, for example, the event of a ball smashing a window. The movement of the ball is of course the producing cause of this event: The transfer of kinetic energy from the ball to the window imparts a physical force (an 'oomph') onto the bonds between the molecules of the glass, causing them to break and the window to shatter. However, there are many other *dependence* causes of the event which were also necessary for the producing cause to have the effect it does (e.g. the tensile strength of the window or the material of the ball). If these conditions were different in some specific way, then the smashing event would not have occurred.

Most physical events are like this, they are brought about by a combination of producing and dependence causes. However, when seeking to (causally) explain an event, we tend to ignore most of its dependence causes and focus primarily on the producing cause(s). That is because, for pragmatic reasons, we are usually interested in identifying only those difference-makers that are most local to and *most specific for* the event in question. Most of the occurrent dependence conditions, such as the tensile strength of a window, are generic, inherent, and familiar properties of the system—and hence they lack sufficient 'causal prominence' (Tseng and Cheng 2024) to be of explanatory value or interest.

The 'driving', mechanistic view of causation in neuroscience assumes, similarly, that only the producing causes in the brain (i.e. the firings of neurons) are going to be explanatorily relevant or causally prominent in the generation of a behaviour. Yet, as we discuss in the next section, this assumption does not fit well with the neurobiology. In the case of the neural causes of behaviour, the dependence conditions in question are not ones that just *happen* to hold, as generic, fixed properties of the neurons involved. They are dynamical, contingent conditions that hold precisely because of the causal influence they exert over whether neurons fire or not; that is their function. Dependence causes in neuroscience are therefore not explanatorily eliminable in the way they often are in the non-biological world. We consider below the many different kinds of dependence conditions that obtain in neural systems, how they come to be established, and how they ultimately support a kind of causal sensitivity in the brain that depends on meaning.

3 | Criterial Causation

Conceptualising the workings of the brain through a driving metaphor that exclusively prioritises 'producing' causes creates the impression that causation between neurons and within neural circuits is fundamentally feedforward, sequential, and deterministic: Neurons get passively driven into action by their presynaptic inputs. But, as most neuroscientists are aware, and as explicitly articulated by Peter Ulric Tse, this is not a complete picture of how neuronal communication works (Tse 2013). Rather, how a neuron responds to incoming activity depends, in large part, on the configuration of its synaptic connections and on other biophysical parameters of the cell (like its current membrane potential). That is, the weights and nature of the synapses between neuron A and neuron B, taken within the context of all of B's other presynaptic inputs, and of the electrophysiological properties of B as a whole, collectively embody what Tse has termed *the neuron's* 'criteria' for *firing*—the conditions that must be met for a neuron to 'release its effect'.

These criteria specify the *types* of presynaptic input the neuron would need to receive in order to produce an action potential (and, by extension, the *types* of input for which the neuron will remain inactive). These can include, for example, a threshold for firing based on number of action potentials arriving over a certain time window. More commonly, however, they specify complex spatiotemporal *patterns* of input to which the neuron is causally sensitive. For example, a neuron, due to its configuration of excitatory and inhibitory synapses, may require a particular *spatial* pattern of inputs for it to 'release its effect', such as those instantiating a logical AND/OR gate. Another neuron might be sensitive to a particular *temporal* pattern, such as a certain rate or timing of inputs.

A neuron's criteria for firing are therefore a type of dependence cause: By changing the criteria (e.g. by changing the weights of its incoming synapses), one can exert control over whether the neuron will fire or not, given the same set of presynaptic inputs. Tse labels this type of causation 'criterial causation'.[2]

Crucially, for our purposes, these criterial dependence causes are not explanatorily eliminable—in the way that dependence conditions in the non-biological world often are—when it comes to understanding neuronal communication and, by extension, how the brain generates behaviour. That is because, first, these criteria are not generic, generalisable properties of neurons. They are contingent, and largely idiosyncratic, features of individual neurons given their specific synaptic configuration and intracellular state. One therefore cannot know whether a postsynaptic neuron will fire based solely on knowledge of its presynaptic action potentials. Second, the conditions placed on a neuron's inputs are dynamic. They are not fixed or static properties of the neuron; they are frequently changing as a result of regular synaptic reconfigurations and the cell's recent firing history. One therefore *also* cannot know whether a postsynaptic neuron will fire based solely on information about its presynaptic action potentials *plus* knowledge of its prior criterial configuration. Given this, some have suggested that 'the state of a neural network might better be described by specifying the state of its synapses than the firing pattern of its neurons. We might even extend this viewpoint by stating that the role of synapses is to control neuronal firing within a neural circuit' (Abbott and Regehr 2004, p.802)—which, we would suggest, is done by specifying the criteria to which neuronal firing is causally sensitive.

Indeed, as Tse has comprehensively argued (2013), and as we will show throughout the remainder of this paper, the ability to change a neuron's criteria through synaptic reconfiguration, sometimes in real time, is ultimately at the heart of how the brain generates behaviour. The configuration and weights of incoming synaptic connections onto any neuron are shaped by the long history of evolution, by learning from individual experience, and by the current state of the organism, including its current cognitive activities. It is these criteria that endow neurons with the functionalities and selective sensitivities that make them useful to the organism.

One might worry, however, that this concept of criterial causation really just refers to situations in which *multiple* different upstream causes are required to *produce* a single downstream effect—and, hence, that the situation is actually entirely compatible with a driving view of causation after all.[3] There is, of course, some sense in which this is true. However, the value of the criterial causation concept is precisely in bringing into focus the role of the dependence relations that are implicitly underlying and, in fact, *creating* such 'many-cause' situations. It draws our attention to the fact that how and why the system came to be configured in such a way that those particular upstream causes bring about that specific downstream effect is fundamental to explaining and understanding even basic neuron-to-neuron communication, let alone how the brain generates behaviour *in toto*.

The neurophysiology of neuronal communication therefore invites us to invert the driving view of causation, in which the activities of some neurons simply drive their downstream partners, given strong enough activation. And to, instead, incorporate the notion of criterial causation into our conceptual toolkit, wherein, due to their sensitivity to *types* of input, downstream neurons ought to be viewed as, in an important sense, *interpreting* the signals they receive (Figure 2). That is, it forces us to consider how and why a neuron came to be configured such that it responds to its inputs in the way that it does.

Note that the presence of this sort of criterial causation within the brain immediately undercuts the intuition that an optogenetically identified neural variable—even when it gives us exogenous control over a given behaviour and is both necessary and sufficient for its occurrence—is *the* explanatorily relevant cause of that behaviour. On the contrary, embracing the concept of criterial causation allows us to see that the identified neural activity can only ever 'drive' a behavioural effect within the context of the rest of the nervous system.

A comprehensive answer to the question 'what is causing this behavior to occur?' must therefore also take into consideration the configuration of the rest of the system. Indeed, as we will discuss below, this concept of criterial causation lays the groundwork for us to see how the configuration of the system grounds the meaning of neural patterns, *which is ultimately what underpins their causal efficacy*. Manipulating these criteria also provides a means of top-down causation by which organisms can alter neural sensitivities, in order to actively guide their own behaviour in real time. This requires a more expansive repertoire of causal concepts, harkening back to the sort of causal pluralism proposed by Aristotle.

4 | Causal Pluralism

Embracing a plurality of causal concepts in neuroscience is therefore essential to developing a full understanding of the causes of behaviour. This is not a novel idea. Aristotle famously developed a scheme incorporating multiple kinds of causation of observed events or phenomena (Falcon 2023). These have been translated as the *material*, *efficient*, *formal*, and *final* causes. These concepts do not map comfortably into modern terms,

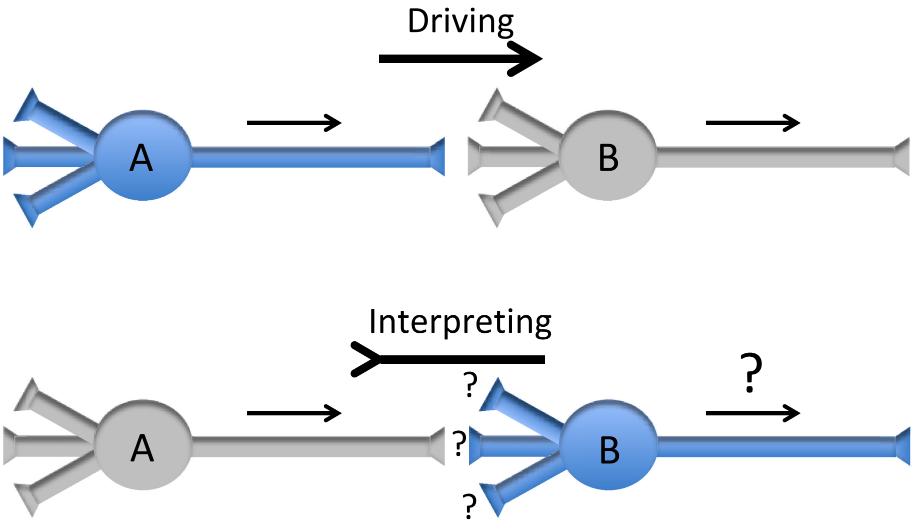

FIGURE 2 | Inverting the driving metaphor. The top row shows a driving relationship between neurons A and B, where B is effectively a passive element—activity in A *drives* activity in B. The bottom row inverts this relationship, highlighting the active role that neuron B plays in *interpreting* its inputs, according to the criteria embodied in its synaptic connections and cellular physiology.

but, very loosely, we can understand the *material* and *efficient causes* as referring to what a thing is made of and what kind of physical forces are acting on it. These concepts therefore roughly capture what might nowadays be called 'mechanism' in neuroscience, and could be seen as referring to the type of synchronic, producing causes that underlie the 'driving' metaphor (Gomez-Marin 2017).

Aristotle's *formal cause* is a somewhat fuzzier concept but is generally taken as referring to the essence or set of properties that makes an object or system that kind of thing and no other; that is, the characteristic way in which the material is organised (its form). For our purposes, the important parallel would be with the *configuration* of the nervous system (including synaptic configurations) and the causal efficacy of the *information* (literally) that configuration represents or embodies (Farnsworth 2022).

Lastly, Aristotle's notion of a *final cause* asks the question: Why did something happen? For what purpose? It thus allows that *having a purpose* can, in its own right, be a cause of something happening. The concepts of formal and final causes are thus essentially diachronic—they reflect the way the system has come to be configured by past events, and the future-directed functionalities that the system enables.

Aristotle therefore took a pluralistic approach to causation, seeing these various causes as complementary ways of explaining natural phenomena, based on different, equally valid perspectives or *types* of causation. Echoes of this way of thinking can also be found in Niko Tinbergen's principles of ethology (Tinbergen 1963), which encompassed four complementary questions, all of which need answering to fully explain a behaviour:

1. Function (or adaption): Why is the animal performing the behaviour?
2. Evolution (or phylogeny): How did the behaviour evolve?
3. Causation (or mechanism): What causes the behaviour to be performed?
4. Development (or ontogeny): How has the behaviour developed during the lifetime of the individual?

Regrettably, in the history of science, Aristotle's formal and final causes were explicitly rejected and talk of organisation or purpose being causally efficacious was largely omitted from polite scientific discourse. Francis Bacon, who was so influential in the 1600's in codifying the scientific method and the *scientific mindset*, argued that science should be solely concerned with material and efficient causes—that is, with mechanism, or matter in motion—i.e. causation as production (Klein 2020). He consigned formal and final causes to metaphysics, or what he called 'magic'.

Yet, as should be clear from the preceding discussion, formal causes—when conceptualised in more modern terms as configuration-based dependence conditions (Farnsworth 2022)—are very much alive and well in systems where criterial causation is at play. One of the key insights from recognising that neurons are not just being passively driven by their presynaptic inputs, but are, in an important sense, *actively* sensing and interpreting these inputs, is that it becomes easy to see how *efficient*—or what we might now call *producing*—causes within the brain (i.e. neuronal firings) necessarily depend on the distribution, organisation, and thus configuration of a given neuron and the system surrounding it (Farnsworth 2018). As we have seen, the organisation of systems, on both a local and global scale, demonstrably helps to set the criteria by which individual neurons will 'release their effect', and is thus undoubtedly a causal factor in the generation of behaviour. Similarly, at a higher level, the criteria for whether a population of neurons will adopt one attractor state or another, given some pattern of inputs, are embodied in the synaptic connections from population A to population B, as well as the connections within population B (Deco and Rolls 2006; Semedo et al. 2019).

It is clear, therefore, that identifying a necessary and sufficient neural mechanism of a behaviour Y, using experimental manipulations, need not imply that one must adopt a horizontally or temporally reductive view with respect to the question of 'what caused Y?'. Instead, our capacity to exogenously control the behaviour shows only that we have learned how to coax the system into behaving in a particular way, by giving it the sort of prompt, stimulus, or information *that it tends to react to in that particular way*. As Alex Gomez-Marin puts it:

> '*We say "circuit X is sufficient" for the cat to behave but what we really mean —and tragically omit — is that "it is sufficient for us to activate circuit X" in order to observe the cat's natural behavior*' (2017, p.6).

In other words, successful optogenetic control requires us to understand only a small amount about *how* behaviour Y is actually being generated, causally speaking—specifically, it relies on knowing only the producing or efficient causes of that effect. For a more comprehensive understanding, we would (at least) need to understand the organisation or configuration of the rest of system, which enables the identified neural activity to have the effect it does.

One might wonder, however, how exactly it is that the system's organisation plays its pivotal causal role? We have argued that it does so by instantiating criterial dependence conditions at the neuronal level, but what exactly does this mean? How should we think of the nature of the causality at play here? This kind of contextual thinking, we suggest, can be understood in terms of the notion of *causal constraints*.

5 | Constraints as Causes

'Constraints are entities, processes, events, relations, or conditions that raise or lower barriers to energy flow without directly transferring kinetic energy'. (Juarrero 2023, p. 49). Consider, for example, a riverbank structuring the flow of water. Or, more relevantly, the configuration of synapses structuring the flow of neurotransmitters and ions between neurons.

The idea that such constraints can act as *causes* may seem controversial if one takes physical forces (i.e. efficient or production causes) to be the only *real* type of causation. However, as we already seen—and as forcefully argued by Alicia Juarrero (1999, 2023), Lauren Ross (2023), Terrance Deacon (2011) and others—causation by constraint is ubiquitous and need not be considered metaphysically problematic: Any structure or process that 'change[s] the dynamics of the underlying processes without being altered themselves (at least not at the same time scale)' (Roli et al. 2022, p.4) can rightfully be thought of as a cause of downstream effects, even without *itself* imparting any 'causal oomph', in virtue of the fact that if one were to intervene on the constraining structure or process in a controlled way, it would lead reliably to changes in the downstream effect. This really is nothing more than saying that the way a system is configured—what physicists call the 'initial conditions' or 'boundary conditions'—will constrain the distribution of physical forces and affect how they play out. And it is in this way, we suggest, that a neuron's 'criteria' for firing gets set: The configuration of the system embodies a set of constraints that structure the flow of energy into a postsynaptic neuron in such a way that sets conditions on the *types* or *patterns* of presynaptic actions potentials to which the postsynaptic neuron will be causally sensitive.

It might be assumed that constraints can only ever be 'limiting' factors: structural features of the system that merely *restrict* the way that energy (or information or causal influence) flows through it (Ross 2023), and therefore do not help to generate or bring about any interesting effects themselves. However, as Juarrero and others have argued, constraints at one level often act as 'enabling' factors, allowing the emergence of functionalities at higher levels (Hooker 2013; Juarrero 1999; Juarrero 2023; Raja and Anderson 2021; Ross et al. 2024; García-Valdecasas and Deacon 2024). To see this, we have to go beyond a simple snapshot perspective on the system, and recognise that the way in which the organisation of a system constrains the possibility space of what happens within it actually *enables* certain phenomena to occur, certain tasks to be performed, and certain (emergent) global or macroscopic properties to obtain, that would otherwise be impossible. As described by Winning and Bechtel: 'By restricting some degrees of freedom of its components and thereby enabling the whole mechanism to do things that would otherwise not be possible, *constraints determine the causal powers of a machine or mechanism*' (Winning and Bechtel 2018, p. 307, our emphasis in bold).

This idea is commonplace in the design of our artefacts. A computer, for example, is designed in such a way as to constrain the flow of electrons within its circuits, to support some functionality. These design constraints do not violate any of the low-level laws of physics—They simply add another level of causation, one that is every bit as important in determining how the system actually behaves. The same is true in living systems—The functionalities that interest biologists, from molecular and cellular to physiological levels, are embodied by sets of constraints (Mitchell 2023a).

These kinds of enabling constraints are ubiquitous in neural systems, where they affect the flow of information and causality, more than energy, per se (though they do have real physical effects on the flow of ions in and out of neurons, rather than directly between them). For example, in the Hodgkin-Huxley model of neuronal conductances, global parameters such as voltage across the membrane affect local variables such as ion channel opening, in turn changing the global electrical field, which feeds back onto the channels, and so on (Hodgkin and Huxley 1952). And as we have seen above, variation in these parameters of cellular excitability, along with those of synaptic transmission, can set the criteria that determine whether or not a downstream neuron will 'release its effect' in response to any given pattern of incoming synaptic activity.

In populations of neurons, we also see *global constraint regimes*, which can generate self-organising dynamics and emergent behaviour, often referred to as '*whole-part causation*'.

The central idea here is that local interactions among parts collectively generate global dynamical *structures* or *fields*, and the order parameters of these global structures can then influence and constrain how the parts behave (Ismael 2011; Juarrero 2023; Prigogine and Stengers 1984). In neural systems, there is good evidence, for example, that collective electrical fields (of the kinds we can detect as local field potentials or by EEG or MEG) can affect individual neuronal excitability. This kind of ephaptic coupling has been proposed as a global control mechanism that can help coordinate neuronal activity (Pinotsis et al. 2023; van Bree et al. 2024). Oscillations of electrical potential are thought to play a similar role, enabling selective communication across brain areas, allowing multiplexed signal transmission, and entraining the timing of neural firings with perceptual or behavioural variables of interest (Buzsáki 2006; Lee et al. 2024; van Bree et al. 2024).

Another kind of self-organising dynamic is evident in the global states and trajectories of activity observed in neuronal populations. The network of excitatory and inhibitory interactions in any interconnected population will lead to the emergence of *attractor states*—i.e. patterns of activity that are more stable and in which the system spends more time (Durstewitz et al. 2023; Ebitz and Hayden 2021; Miller 2016). These states thus reflect the way in which the global *constraint regime*, embodied in the organisation of the network, can be said to be an enabling cause of phenomena such as low-dimensional manifolds, in virtue of the way in which it constrains the possibility space of activity within the network (Silberstein and Chemero 2013; Ross et al. 2024).

In these systems, it should hopefully be clear that the arrow of causation is not exclusively bottom-up: It is not just neuronal cause-and-effect 'driving' behavioural outcomes. The constraint regimes responsible for setting a neuron's criteria for firing, and for enacting different forms of whole-part causation and attractor states, are *also* essential causal contributors to how the brain is generating behaviour. And thus are necessary to understand in order to *explain* behavioural phenomena (Durstewitz et al. 2023; Gallego et al. 2017; Robson and Li 2022).

We therefore argue that the organisation of the system and the dynamical constraint regime it embodies are a key part of the causal story of any given behaviour, and are therefore in need of explanation if we are to fully understand how behaviour is being generated. This means we have to look beyond mechanistic and synchronic 'how' questions and also ask diachronic 'why' questions to fully explain behaviour (Marr and Poggio 1976; Tinbergen 1963). First, 'why' in the sense of 'how come?'—How did the system come to be organised in such a way that it embodies the particular constraint regime it does? And, second, 'why' in the sense of 'what for?'—What is the reason for the system to be organised in this way rather than another way? Taking the 'how come' question first, philosopher Fred Dretske has argued that this speaks to another type of cause relevant to the question of '*what causes a behavior to occur?*' He refers to this as a structuring cause.

6 | Structuring Causes and Final Causes

In his account of mental causation, Dretske introduces a helpful distinction between *triggering causes* and *structuring causes* of behaviour (Dretske 1988). A triggering cause is an event, stimulus or condition that initiates the process that ultimately leads to the performance of, for example, a mouse's feeding behaviour. A triggering cause could therefore be the onset of a food stimulus. Similarly, the triggering cause of a car engine starting could be the turning of a key in the ignition.

A structuring cause, on the other hand, is an event that helps to create or shape *the process* itself; that is, the process that gets initiated by the triggering cause and that leads to the execution of the behaviour in question. A structuring cause could therefore be the wiring of a car or the event(s) that help to shape the neurophysiology of the mouse. As Dretske put it:

> 'In looking for the cause of a process, we are sometimes looking for the triggering event: what caused the event C which caused the M [the behavioral phenomenon]. At other times we are looking for the events that shaped or structured the process: *what caused C to cause M rather than something else*. The first type of cause, the triggering cause, causes the process to occur now. The second type of cause, the structuring cause, is responsible for its being this process, one having M as its product, that occurs now'. (1988, pp. 42–45; our emphasis).

Structuring causes are what enable the triggering cause to have the observable behavioural effect it does. In other words, structuring causes *cause* the constraint regime embodied in the system's organisation. These are distal (i.e. historical) events or conditions, over both evolutionary and individual timescales, that are thus every bit as much a part of the causation of a behaviour as the currently active neural states. This view aligns well with the perspectives of *process philosophy*, wherein living organisms are to be seen as temporally extended processes, rather than objects or substances whose existence can be captured in instantaneous states (Meincke 2019; Nicholson and Dupré 2018: Seibt 2016).

In addition to the 'how come?' question, we can also ask the 'what for?' question. The current organisational structure of any living system reflects the evolutionary history of the organism and is, thus, necessarily oriented towards a function or *purpose* to persist (Ellis 2012, 2016; Mitchell 2023a)—in the sense that, in most cases, a system's macroscale organisation has been selected for *because* it helps to constrain microscale activity in a way that both enables and promotes survival-enhancing behaviour. Indeed, this was Aristotle's insight with his fourth type of cause, the *final cause*. He thought that a defining characteristic of animal behaviour was its purposive and goal-directed nature, and that this needed to be recognised in the causal schema one uses to understand and explain behaviour.

Talk of final causes and organismal purposiveness can appear somewhat vague a3nd perhaps even magical. However, we contend that Dretske's work on structuring causes helps to operationalise it in concrete terms. In particular, Dretske emphasises the role of learning and experience in shaping the neurophysiology of an organism. In the language of constraints, this means that the personal history of the organism causes changes to the global constraint regime, thereby acting as a structuring cause of its subsequent behaviours (for the reasons given above) in a way that is entirely natural and non-mysterious. Likewise, the idea of a final cause, within this framework, does not need to entail some kind of retrocausality, with a future state reaching back in time to influence current behaviour; it is simply *the current possession of a goal state* (towards a desired future end) that has causal power in the system.

What this means is that one of the main causes of an organism's behaviour is quite literally its own historical interactions with the world and its past experiences. As we will argue in the coming sections, these interactions essentially build meaning and subjectivity into the causal architecture of the system, which is what ultimately guides its behaviour. If one buys this argument, then it becomes clear that if we want a full understanding of the causes of behaviour, we need to understand both the historical processes that allow organisms to acquire reasons and the current processes that enable them to act on the basis of those reasons.

As Michael Silberstein puts it: 'For any particular synchronic-frame or still-shot of a biological system at a time t with some duration d, the determining features include diachronic multiscale interactions (context sensitivity) and global constraints outside the time-slice in question'. He goes on to argue that: 'when it comes to such complex biological systems one should take the word process very seriously and understand that such systems are spatially, temporally, functionally and in a thin sense teleologically extended'. (Silberstein 2021, pp. 370–371).

7 | Macroscopic Causation and Informational Causation

Moving beyond a 'driving' conception of causation within the brain, and embracing criterial causation, creates the conceptual space necessary to see how macroscopic causes can exist within the brain. We have already seen how it enables global variables, dynamics, and constraints to be causally efficacious within living systems by virtue of setting (or *structuring*) the causal sensitivities of neurons and neural populations. However, it is also crucial to reiterate that, in most cases, this means that individual neurons or populations of neurons are tuned to respond to macroscopic *patterns* (i.e. spatiotemporally extended *types*) of incoming activity, rather than the specific details. This is true, for example, for neurons that respond to the *rate* of inputs over some time window, but which do not distinguish temporal patterns within such windows. And it is true for populations of neurons that are selectively responsive to low-dimensional (macroscopic) patterns in their inputs, rather than the high-dimensional (microscopic) details of each individual presynaptic neuron's firing (Ebitz and Hayden 2021; Gallego et al. 2017; Semedo et al. 2019). In the population-coding paradigm, it is these higher-order patterns that are thought to carry causal weight within the system (Barack and Krakauer 2021; Ebitz and Hayden 2021; Mitchell 2023a, 2023b; Semedo et al. 2020). This also aligns with the important observation that the *lack of firing* of given neurons can be just as causally effective in the system as the firing of neurons (e.g. Pérez-Ortega et al. 2024).

This kind of sensitivity to macroscopic patterns is observed empirically and is consistent with theoretical work demonstrating the efficacy of *macroscopic causation* (Comolatti and Hoel 2022; Ellis 2012; Flack 2017; Hoel et al. 2013; Rosas et al. 2024). In this view, what happens in the system is sensitive to the macrostates that subsystems within it occupy (and how they are interpreted by other subsystems), rather than the details of the microstates by which they are transiently realised. This is broadly akin to the way in which, in language, we are generally sensitive to the word that is being uttered, rather than to the specific acoustic and prosodic features of *how* it was uttered on that specific occasion.

Of course, any given macrostate must always be instantiated by some specific microstate at a given moment and one could argue that is where the real causation lies—at the lowest level of physical detail. However, two considerations argue against this interpretation.

First, due to the inherent noisiness of neuronal signalling and molecular and cellular processes in general (Faisal et al. 2005; Glimcher 2005; Rusakov et al. 2020; Sanborn et al. 2024), the microscale details of the system at any moment will not fully determine (in the sense of causally *necessitating*) what happens next (Tse 2013). This does not mean that the outcome will necessarily be settled by *some particular* random jigglings or jitterings at the molecular scale, however. What it does, in the words of physicist George Ellis, is introduce some *causal slack* into the system (Ellis 2008, 2012; Mitchell 2023a). This means—as we have seen in Sections 5 and 6—that the organisation of the system can come to embody some higher-order constraints that really do have causal efficacy over how the system evolves (because the causation is not already exhausted by the lower-level details (Kim 1993)).

Second, and as a result of this causal slack, these systems also come to be sensitive to higher-order patterns, or macrostates that are *multiply realisable*—that is, where any given macrostate may be realised by many different microstates that are causally equivalent within the system (Figure 3). If one understands causation in a *counterfactual sense* (List and Menzies 2017; Menzies and List 2010; Sinnott-Armstrong 2019; Woodward 2005), then what this means is that the causal sensitivity of the system, in these cases, lies at the level of these coarse-grained patterns, rather than the details of their neuronal instantiations (Albantakis et al. 2019; Semedo et al. 2019). That is, many changes to the microstate will *not* affect the outcome, unless they *also* change the coarse-grained macrostate in a way that the downstream neurons are sensitive to. Crucially, because they average, spatially and temporally, over microscopic noise, the coarse-grained macrostate patterns contain more *effective information* about the future (and past) states of the system than any given detailed and momentary microstate (Hoel et al. 2013; Rosas et al. 2024). Indeed, this must be the case for robust signalling in a network with noisy components (Deco et al. 2009; Tsimring 2014).

In this sense, then, the system is causally sensitive to patterns or *types* of activity, constituted by *equivalence classes* of microscale details, *which are established by* the criterial configuration of downstream neuron(s) (Buzsáki 2010; Tse 2013). In other words, *it is the configuration of the neurons interpreting the signals*

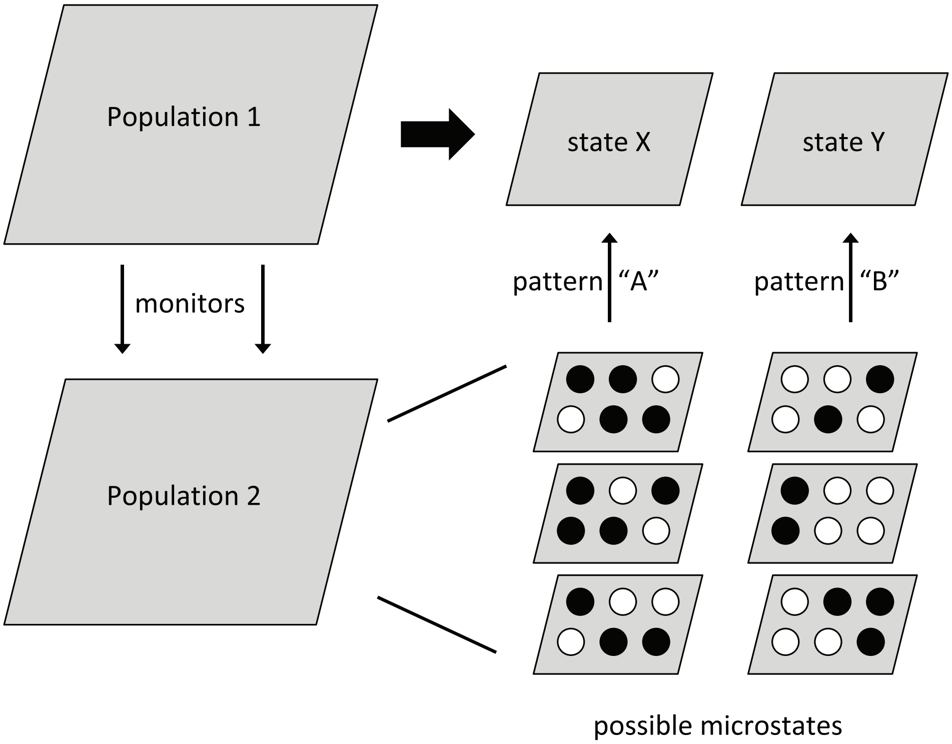

FIGURE 3 | Multiple realisability and macroscopic causation. A given population of neurons (Population 1) will monitor its inputs, depending on the criteria embodied in its afferent and internal connectivity. These criteria will determine the causal sensitivity and response to incoming macroscopic patterns (A versus B), which are each realisable in multiple possible microstates (reprinted, with permission, from Mitchell 2023a).

which generates the equivalence classes, by virtue of their sensitivity to patterns and insensitivity to details – effectively, a filtering or categorisation of their inputs. This is a form of, what we would call *informational causation*: The system becomes causally sensitive to information that is, to a large extent, *created* by the downstream neuron(s), not simply received or transmitted to them. The meaningful information in the system (i.e. what counts as 'signal'/pattern) is not inherent in the presynaptic inputs themselves; it inheres in the active and selective *interpretation* of those inputs.

As we have seen, the way that these downstream neuron(s) come to interpret their inputs reflects, at least in part, the prior influences that have configured the system with the sensitivities it has. This raises the question: why do these prior influences affect the system in the way that they do? That is, why do the experiences of the system lead it to exhibit the sort of informational economy it does, and not a different sort of informational economy? The answer, we suggest, is that such historicity builds meaning and subjectivity into the causal architecture of the system by tuning neurons and neural populations to be sensitive to *semantic information*, making the macroscopic causation within the brain a form of *meaningful causation*.

8 | Pragmatic and Semantic Meaning

We have already seen how the informational economy (embodied in the physical, dynamical configurations of the system, the constraint regime it enacts, and the neuronal 'criteria' this creates) is shaped by the system's historicity, such that it comes to reflect or instantiate the subjective perspective of the organism itself. In this section, we argue that this causally efficacious 'subjective perspective of the organism' should be viewed as a realisation of what is meaningful to the organism (Jaeger et al. 2024).

There are two senses in which patterns of neural activity can be meaningful for an organism: First, they may be *about something*. And second, they may be *for something*. The aboutness is most obvious for perceptual states, which typically reflect the presence of some stimulus in the environment at a current moment. More precisely, they represent an *inference* or belief about the existence of some objects out in the world that are the causes or sources of the incoming sensory data. An internal pattern can usefully represent such an object by virtue of 'standing in exploitable relation to it' (Shea 2018). That is, having such an internal representation allows the organism to take some action in relation to the object, which it could not do otherwise. This relates to the second criterion—that such internal representations be *useful* for something, where the usefulness depends on their 'content' (Millikan 1984, 1995). This links to the second sense of meaning, which is not just of aboutness, but salience or value to the organism. Such internal representations are not just *referential*, they are also, potentially at least, *consequential*.

In the simplest cases, the organism may have preconfigured control policies, which directly induce behavioural responses to particular stimuli. For example, lamprey will move away from a large, looming shadow (a potential predator), but towards a small, moving object (potential prey) in their visual field (Cisek 2019). Many species have similar prewired escape circuits and other innate approach/avoid preferences. We may say in these cases that the meaning is *pragmatic*—It is baked into the adaptiveness of the responses to the various stimuli (Mitchell 2023b). A purely synchronic explanation would locate the causation in the neural mechanisms of stimulus detection and linked action, but there is clearly also a kind of diachronic causation at play in the (evolutionary and developmental) *structuring* causes that led the system to be so configured.

In more complex cases, perceptual systems generate internal, genuinely *semantic* representations, which are decoupled from obligate action, and which are simply reported or made available to other parts of the nervous system (Mitchell 2023b). We call these semantic because they are *indicative*, rather than *imperative*. The meaning of these internal patterns of neural activity is grounded through the organism's individual history of sensorimotor exploration (Bahrick and Lickliter 2002; Barsalou 2008; Gopnik and Wellman 2012; Pezzulo and Castelfranchi 2007). This builds up a stored context of useful knowledge—about objects, their properties, their causal relations to other things, and their affordances for the organism.

The meaning of such states is thus not in the isolated, active states themselves, but is relational and distributed through the web of synaptic connections that embodies these kinds of knowledge (Barsalou 2008; Blouw et al. 2016). This kind of view can thus reconcile computational theories of mind (which involve operations over currently active states comprising 'symbolic representations') with connectionist theories (which supply the stable background context that grounds the meaning of the active states) (Mitchell 2023b; Piccinini 2022).

If we want to understand the causes of an organism's behaviour at any moment, we thus need to consider what its internal representations are and what those representations *mean* to the organism, based on its history, stored (distributed) knowledge, and prewired or learned control policies (embodying pragmatic or semantic meaning). Because of coarse graining and multiple realisability, the causation in this kind of system cannot be explicated entirely in terms of the active, synchronic neural mechanisms, the details of which are often arbitrary and incidental (Menzies and List 2010; Rosas et al. 2024) and which can even drift over time (Rule et al. 2019; Driscoll et al. 2022). Instead, it derives primarily from the *meaning* of these internal states—what the organism believes about the world and the threats and opportunities it presents—which results from its experiences through time.

This view was well articulated by Walter Freeman, who, several decades ago, anticipated the now-popular 'population doctrine' of neural coding and the action-oriented and affordance-laden nature of neural representations: [These patterns of neural activity] 'do not represent external objects; they embody and implement the meanings of objects for each individual, in terms of what they portend for the future of that individual, and what that individual should do with and about them' (Freeman 2000, p.93). Modern systems neuroscience is now reinforcing this meaning-laden view of the global patterns of neural dynamics (e.g. Thura et al. 2022; González-Rueda et al. 2024; Khilkevich et al. 2024; Zutshi et al. 2024).

Of course, what an organism chooses to actually do in any given situation will also reflect its current internal states and motivational needs, as well as any ongoing goals or plans. The criteria for action are thus changing all the time and the system can be reconfigured on the fly to reflect this. A neuron or neural population's 'criteria' are therefore changeable, over slow timescales by learning, but also over very rapid timescales, in response to the very recent history of firing and incoming signals, including neuromodulators (Tse 2013). Synaptic weights between neurons are constantly being reconfigured on millisecond timescales, by contextual signals, which alter the gain and change the sensitivities to various incoming patterns. These can include effects of attention, arousal, oscillatory entrainment of the type described above, top-down expectations, the selection of goals, and so on (Dayan 2012; Shine et al. 2021; Shine 2023; Taylor et al. 2024; Thiele and Bellgrove 2018).

These kinds of control mechanisms can be taken as examples of *top-down causation* (Ellis 2009), in two senses. First, in a functional sense of information flowing from brain regions comprising higher levels of the functional hierarchy (concerned with the adoption and prioritisation of goals, for example) and constraining the dynamics of regions comprising lower levels (e.g. those concerned with shorter term action selection). And second, in a more controversial ontological sense of causation flowing from an emergent 'mental' (or even just cognitive) level to the neural levels 'below'—i.e. in a way that depends on the meaning or content of mental states (discussed more below).

9 | Summary

So where does this leave us with respect to our original problem of how to conceptualise the causes of a behaviour, especially in light of our newfound ability to exogenously control certain behaviours through direct manipulation of neuronal activity?

We suggest that, equipped with the full suite of more expansive causal concepts argued for above, it is clear that even 'necessary and sufficient' synchronic neural variables, which give us the capacity to exogenously control a particular behaviour, are only ever a very small part of the story of what causes a behaviour to occur. In particular, they are the 'triggering cause' of the behaviour. To fully understand, and thus explain, the occurrence of any given behaviour we *also* need to consider (i) the constraint regime that the neural mechanism is situated within, (ii) the nature of the informational economy that constraint regime enacts (i.e. the macroscopic *pattern* that downstream neurons are causally sensitive to, and that the identified neural variable forms a part of), and (iii) the structuring causes of all of this (i.e. the historicity of the system). Collectively, this would give us an insight into the organism's meaningful, subjective perspective, embodied within this informational economy.

Each of these types of non-reductive, diachronic forms of causation is important to consider if one is to properly understand *why* manipulation of that neural mechanism enables control over the behaviour. Overall, this paints a very different picture to the horizontally, vertically, and temporally reductive view of causation implied by the dominant (if often implicit) 'driving' metaphor.

10 | Discussion

The foregoing discussion brings us to two categories of causation that have been deemed for centuries by many philosophers and scientists to be metaphysically problematic: mental causation and agent causation.

10.1 | Mental Causation

René Descartes famously proposed a distinction between physical stuff and mental stuff. This 'substance dualism' allowed him to privilege goings-on in the mental realm and protect them from reduction to the merely physical (Robinson 2023). The problem with this scheme, as pointed out by his correspondent Elisabeth, Princess of Bohemia, is it left no way for mental goings-on to influence things in the physical realm. How could the abstract content of an immaterial thought push physical things around in the brain in the way that it must in order to have any causal efficacy?

The discussion above, which extends views of causation beyond the synchronic and mechanistic, offers a way to reconceptualise this problem. Thoughts are not immaterial. They are *patterns of meaningful neural activity.* They can thus—unproblematically—have physical causal efficacy in the neural system in the normal way that patterns of activity do, but this will crucially be conditioned on what they mean. This meaning is grounded by experience and interpreted through the distributed network of synaptic connections embodying stored knowledge and control policies—i.e. the criteria described above (Mitchell 2023a) (Figure 4).

The notion that abstract things, like concepts or ideas, could affect physical things, like neurons, or the flow of ions, has an obvious parallel in computing, where the concepts encoded in software constrain the physical workings of the computer (Ellis 2016; Rosas et al. 2024). Computer scientist Subrata Dasgupta has called this 'liminal causation'—that is, causation that acts at the border of the abstract and the concrete (Dasgupta 2016). The steps of an algorithm—itself an abstract object that could be realised in many different physical systems—become realised in hardware through the actions of compilers, assemblers, and other elements of an operating system that ultimately constrain the flow of electrons through the transistors of the computer.

Similarly, the abstract content of our thoughts—percepts, beliefs, desires, intentions, and so on, that are about specific things—will have causal efficacy in the system depending on that content (Ellis 2016; Mitchell 2023a). Organisms can figure out what to do, in a way that is genuinely causally based on the semantic content of their beliefs and desires, and not just on their neural vehicles.

10.2 | Agent Causation

As neuroscience elucidates more and more details of the neural mechanisms underlying behaviour, we have at least two ways of interpreting these discoveries. We could see them as showing us *the means by which organisms regulate their behaviour.* Or we could see them as identifying *the neural causes of behaviour* (or, as Barack et al., say: 'the events in the brain that "cause" behavior' (Barack et al. 2022)). Under the latter framing, the organism, as an agent, disappears from our

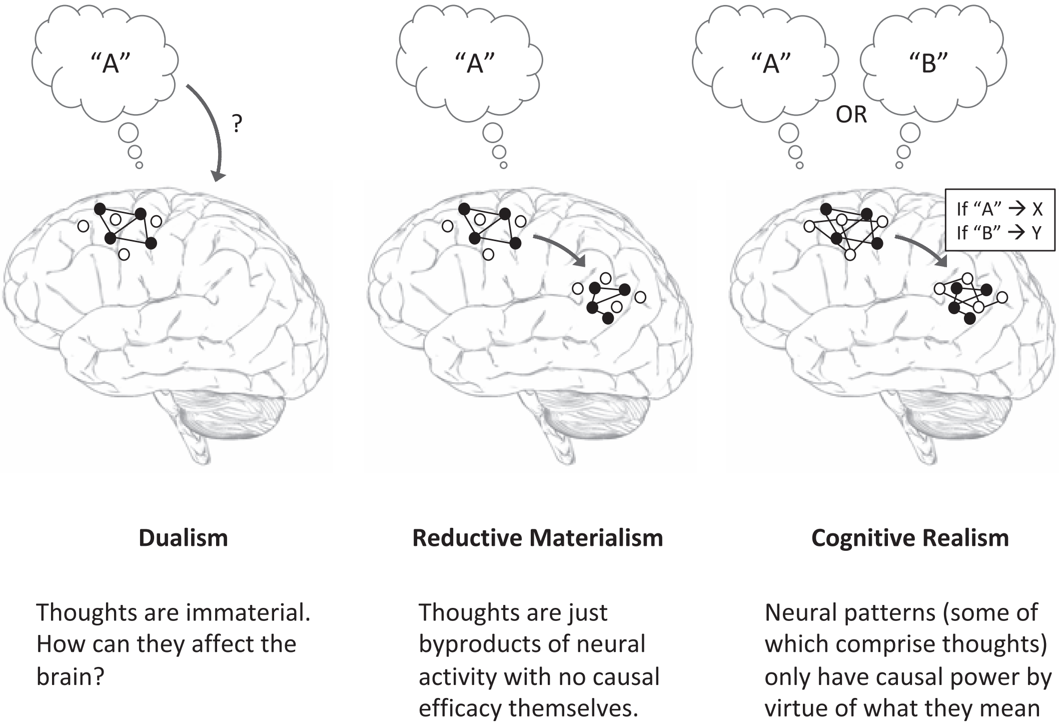
"A"
?
"A"
"A"
OR
"B"
If "A" → X
If "B" → Y
Dualism
Reductive Materialism
Cognitive Realism
Thoughts are immaterial. How can they affect the brain?
Thoughts are just byproducts of neural activity with no causal efficacy themselves.
Neural patterns (some of which comprise thoughts) only have causal power by virtue of what they mean

FIGURE 4 | Mental causation. Under a dualistic viewpoint, thoughts are somehow non-physical, making it a mystery how they could affect neural processes. Under reductive materialism, thoughts are epiphenomenal; all the causal work is done by their neural vehicles. Under what can be called 'cognitive realism' (Mitchell 2023a), thoughts are meaningful patterns of neural activity that have causal efficacy in the system based on their meaning (reprinted, with permission, from Mitchell 2023a).

explanations of its own behaviour (Franklin 2014). It is not really doing anything or deciding anything—It is not the cause of anything. It is just being pushed around by events that are happening within it.

This view—of *event causation*—has been popular in philosophy for some time (Davidson 1963; Franklin 2018), yet it relies on a reductive and synchronic perspective which, we argue, is not congruent with the true nature of living organisms. The importance of a temporally extended, diachronic view of causation has been emphasised above, as well as the dangers of 'vertically' reducing cognitive states and operations to their neural realisers, both of which serve to strip *meaning* out of the system. But there is also a danger of what we have called 'horizontal reductionism' or *causal isolationism*, which identifies the causes of behaviour with just some discrete and localisable subset of neural states within the whole nervous system (Potter and Mitchell 2022).

This horizontally reductive view comes naturally from the types of experiments that we perform. To generate robust paradigms of behaviour, we typically restrict the choices of the organism to binary options, control as many variables as we can, remove all possible contextual factors, exhaustively train the animal on the task, where it has nothing else it needs to care about, and then focus on some particular neural region where activity patterns or levels correlate with one outcome or the other. We then take activity in that region to be 'the cause of the behavior'. This conclusion can be reinforced by experiments that strongly drive activity in that area using optogenetic techniques in a way that directly brings about the behaviour.

The problem, of course, is that no brain region or circuit does anything in isolation (Gomez-Marin 2017; Pessoa 2022). Nor does nature present itself so obligingly to animals in the real world—one stimulus at a time, one task at a time. Living organisms—as agents—have to actively manage their own behaviour over nested timescales, balancing and prioritising multiple needs and goals, sustaining ongoing plans and activities, while adapting to changing circumstances and accommodating to new information, in order to navigate complex and dynamically varying environments, typically while coping with interference from other agents with their own varying goals. Making an all-things-considered judgement about what best to do in any given scenario requires input from subsystems *across the whole brain* (Ismael 2016; Mitchell 2023a).

There will thus be a multiplicity of causal factors feeding into any behaviour, in a non-decomposable way, through a web of contextual conditionalities. These are processed by distributed circuits and brain regions, but integrated for the purpose of holistic decision-making. One way to think of this is as a collective, massively parallel optimisation problem, with each area trying to satisfy (or 'satisfice') its own constraints, based on the 'criteria' instantiated in its connections, and the current incoming data and modulatory influences from other areas (Pessoa 2022; Robson and Li 2022; Suzuki et al. 2023), until a global consensus (or lowest energy state) emerges.

We contend that this *just is* the agent deciding what to do, for its own, agent-level reasons, as best it can with the information and neural resources and time that it has. There will of course be some particular neural mechanisms that an organism *is using* to carry out these operations, but, rather than identifying those mechanisms as *the causes* of the behaviour, it seems valid and appropriate, given the ineliminable contextual and historical dependencies at play, to say that a behaviour occurred because an organism decided to do it. From this perspective, we can see that the organism itself is the appropriate *locus of causation* of its behaviour, as a holistic entity with continuity through time (Potter and Mitchell 2022; Mitchell 2023a)—It is a causal agent unto itself. And our explanations ought to reflect that.

11 | Conclusion

The amazing progress of neuroscience in recent years is something of a double-edged sword. On one hand, it offers unprecedented power to causally intervene in the system, activating neural mechanisms that appear to 'drive' all kinds of interesting and important behaviours. On the other hand, it threatens to reduce our understanding of behaviour and agency to *nothing more than* synchronic neural mechanisms.

We have argued that a full explanation of the nature of the causes of behaviour requires the dimension of time. A purely synchronic view of neural mechanisms misses out on the very property that defines living beings: *historicity*. Living beings are historical processes, with extension in time. They accumulate causal power by accumulating causal knowledge of the world, using it to guide and manage their behaviour in an integrative and holistic fashion. We have outlined here some of the philosophical resources that neuroscientists can draw on to enrich our notions of causation to reflect these diachronic and non-reductive features of life.

Author Contributions

Both authors contributed equally to conceptualisation and writing.

Acknowledgements

HP was supported by a Provost's Fund grant (1481.9050961) from Trinity College Dublin.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

We do not present any data.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ejn.70064.

Endnotes

[1] According to the Cambridge dictionary (n.d.), to 'drive' is 'to force someone or something into a particular state' or 'to force someone or something to go somewhere or do something'.

[2] It should be noted that Tse (2013) uses this term to refer, both, to the causal *effects* of a neuron's criteria (i.e. the criterial dependence cause) and to the *causing* of the criteria itself (i.e. the events that *create* the

dependence conditions themselves). For parsimony reasons, we will use the term to refer to the former definition only.

[3] We thank an anonymous reviewer for pressing us to clarify this point.

References

Abbott, L. F., and W. G. Regehr. 2004. "Synaptic Computation." *Nature* 431, no. 7010: 796–803.

Albantakis, L., W. Marshall, E. Hoel, and G. Tononi. 2019. "What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks." *Entropy* 21, no. 5: 459.

Bahrick, L. E., and R. Lickliter. 2002. "Intersensory Redundancy Guides Early Perceptual and Cognitive Development." *Advances in Child Development and Behavior* 30: 153–187.

Barack, D. L., and J. W. Krakauer. 2021. "Two Views on the Cognitive Brain." *Nature Reviews Neuroscience* 22, no. 6: 359–371.

Barack, D. L., E. K. Miller, C. I. Moore, et al. 2022. "A Call for More Clarity Around Causality in Neuroscience." *Trends in Neurosciences* 45, no. 9: 654–655.

Barsalou, L. W. 2008. "Grounded Cognition." *Annual Review of Psychology* 59: 617–645.

Blouw, P., E. Solodkin, P. Thagard, and C. Eliasmith. 2016. "Concepts as Semantic Pointers: A Framework and Computational Model." *Cognitive Science* 40, no. 5: 1128–1162.

Boyden, E. S., F. Zhang, E. Bamberg, G. Nagel, and K. Deisseroth. 2005. "Millisecond-Timescale, Genetically Targeted Optical Control of Neural Activity." *Nature Neuroscience* 8, no. 9: 1263–1268.

Brembs, B. 2021. "The Brain as a Dynamically Active Organ." *Biochemical and Biophysical Research Communications* 564: 55–69.

Buzsáki, G. 2006. *Rhythms of the Brain*. Oxford University Press.

Buzsáki, G. 2010. "Neural Syntax: Cell Assemblies, Synapsembles, and Readers." *Neuron* 68: 362–385.

Buzsáki, G. 2019. *The Brain From Inside Out*. Oxford University Press.

Cambridge. n.d. "Drive." In cambridge.org Dictionary. Retrieved November 21, 2024, from https://dictionary.cambridge.org/dictionary/english/drive.

Castaneda, A. N., A. Huda, I. B. Whitaker, et al. 2024. "Functional Labeling of Individualized Postsynaptic Neurons Using Optogenetics and Trans-Tango in Drosophila (FLIPSOT)." *PLoS Genetics* 20, no. 3: e1011190.

Cisek, P. 1999. "Beyond the Computer Metaphor: Behaviour as Interaction." *Journal of Consciousness Studies* 6, no. 11–12: 125–142.

Cisek, P. 2019. "Resynthesizing Behavior Through Phylogenetic Refinement." *Attention, Perception & Psychophysics* 81, no. 7: 2265–2287.

Cobb, M. 2020. *The Idea of the Brain: The Past and Future of Neuroscience*. Hachette UK.

Comolatti, R., and E. Hoel. 2022. "Causal Emergence Is Widespread Across Measures of Causation." arXiv preprint arXiv:2202.01854.

Dasgupta, S. 2016. *Computer Science: A Very Short Introduction*. Oxford University Press.

Davidson, D. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60, no. 23: 685–700.

Dayan, P. 2012. "Twenty-Five Lessons From Computational Neuromodulation." *Neuron* 76, no. 1: 240–256.

Deacon, T. W. 2011. *Incomplete Nature: How Mind Emerged From Matter*. WW Norton & Company.

Deco, G., and E. T. Rolls. 2006. "Decision-Making and Weber's Law: A Neurophysiological Model." *European Journal of Neuroscience* 24: 901–916.

Deco, G., E. T. Rolls, and R. Romo. 2009. "Stochastic Dynamics as a Principle of Brain Function." *Progress in Neurobiology* 88, no. 1: 1–16.

Dewey, J. 1896. "The Reflex arc Concept in Psychology." *Psychological Review* 3, no. 4: 357–370.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of causes*. MIT Press.

Driscoll, L. N., L. Duncker, and C. D. Harvey. 2022. "Representational Drift: Emerging Theories for Continual Learning and Experimental Future Directions." *Current Opinion in Neurobiology* 76: 102609.

Durstewitz, D., G. Koppe, and M. I. Thurm. 2023. "Reconstructing Computational System Dynamics From Neural Data With Recurrent Neural Networks." *Nature Reviews Neuroscience* 24, no. 11: 693–710.

Ebitz, R. B., and B. Y. Hayden. 2021. "The Population Doctrine in Cognitive Neuroscience." *Neuron* 109, no. 19: 3055–3068.

Ellis, G. F. 2008. "On the Nature of Causation in Complex Systems." *Transactions of the Royal Society of South Africa* 63, no. 1: 69–84.

Ellis, G. F. 2009. "Top-Down Causation and the Human Brain." In *Downward Causation and the Neurobiology of Free Will*, edited by N. Murphy, G. Ellis, and T. O'Connor, 63–81. Springer.

Ellis, G. F. 2012. "Top-Down Causation and Emergence: Some Comments on Mechanisms." *Interface Focus* 2, no. 1: 126–140.

Ellis, G. F. 2016. *How can Physics Underlie the Mind?: Top-Down Causation in the Human Context*. Imprint, Springer.

Faisal, A. A., J. A. White, and S. B. Laughlin. 2005. "Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring." *Current Biology* 15, no. 12: 1143–1149.

Falcon, A. 2023. "Aristotle on Causality." In *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*, edited by E. N. Zalta and U. Nodelman. https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/.

Farnsworth, K. D. 2018. "How Organisms Gained Causal Independence and How It Might Be Quantified." *Biology* 7, no. 3: 38.

Farnsworth, K. D. 2022. "How an Information Perspective Helps Overcome the Challenge of Biology to Physics." *Biosystems* 217: 104683.

Filipowicz, A., J. Lalsiamthara, and A. Aballay. 2022. "Dissection of a Sensorimotor Circuit Underlying Pathogen Aversion in C. Elegans." *BMC Biology* 20, no. 1: 229.

Flack, J. C. 2017. "Coarse-Graining as a Downward Causation Mechanism." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375, no. 2109: 20160338.

Franklin, C. E. 2014. "Event-Causal Libertarianism, Functional Reduction, and the Disappearing Agent Argument." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 170, no. 3: 413–432.

Franklin, C. E. 2018. *A Minimal Libertarianism: Free Will and the Promise of Reduction*. Oxford University Press.

Freeman, W. J. 2000. "A Neurobiological Interpretation of Semiotics: Meaning, Representation, and Information." *Information Sciences* 124, no. 1: 93–102.

Gallego, J. A., M. G. Perich, L. E. Miller, and S. A. Solla. 2017. "Neural Manifolds for the Control of Movement." *Neuron* 94, no. 5: 978–984.

García-Valdecasas, M., and T. W. Deacon. 2024. "Origins of Biological Teleology: How Constraints Represent Ends." *Synthese* 204, no. 2: 75.

Glimcher, P. W. 2005. "Indeterminacy in Brain and Behavior." *Annual Review of Psychology* 56: 25–56.

Gomez-Marin, A. 2017. "Causal Circuit Explanations of Behavior: Are Necessity and Sufficiency Necessary and Sufficient?" In *Decoding*

Neural Circuit Structure and Function: Cellular Dissection Using Genetic Model Organisms, edited by A. Çelik and M. F. Wernet, 283–306. Springer International Publishing.

González-Rueda, A., K. Jensen, M. Noormandipour, et al. 2024. “Kinetic Features Dictate Sensorimotor Alignment in the Superior Colliculus.” *Nature* 631, no. 8020: 378–385.

Gopnik, A., and H. M. Wellman. 2012. “Reconstructing Constructivism: Causal Models, Bayesian Learning Mechanisms, and the Theory Theory.” *Psychological Bulletin* 138, no. 6: 1085–1108.

Hall, N. 2004. “Two Concepts of Causation.” In *Causation and Counterfactuals*, edited by J. Collins et al., 225–276. MIT Press.

Hodgkin, A. L., and A. F. Huxley. 1952. “A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve.” *Journal of Physiology* 117, no. 4: 500–544.

Hoel, E. P., L. Albantakis, and G. Tononi. 2013. “Quantifying Causal Emergence Shows That Macro can Beat micro.” *Proceedings of the National Academy of Sciences of the United States of America* 110, no. 49: 19790–19795.

Hooker, C. 2013. “On the Import of Constraints in Complex Dynamical Systems.” *Foundations of Science* 18: 757–780.

Ismael, J. 2016. *How Physics Makes Us Free*. Oxford University Press.

Ismael, J. T. 2011. “Self-Organization and Self-Governance.” *Philosophy of the Social Sciences* 41, no. 3: 327–351.

Jaeger, J., A. Riedl, A. Djedovic, J. Vervaeke, and D. Walsh. 2024. “Naturalizing Relevance Realization: Why Agency and Cognition Are Fundamentally Not Computational. Hypothesis and Theory.” *Frontiers in Psychology* 15: 1362658.

James, W. 1890. *The Principles of Psychology*. Vol. 1. Henry Holt and Co.

Juarrero, A. 1999. *Dynamics in Action: Intentional Behavior as a Complex System*. MIT Press.

Juarrero, A. 2023. *Context Changes Everything: How Constraints Create Coherence*. MIT Press.

Khilkevich, A., M. Lohse, R. Low, et al. 2024. “Brain-Wide Dynamics Linking Sensation to Action During Decision-Making.” *Nature* 634, no. 8035: 890–900.

Kim, C. K., A. Adhikari, and K. Deisseroth. 2017. “Integration of Optogenetics With Complementary Methodologies in Systems Neuroscience.” *Nature Reviews Neuroscience* 18, no. 4: 222–235.

Kim, J. 1993. “The Myth of Nonreductive Materialism.” In *Supervenience and Mind*, edited by J. Kim, 265–284. Cambridge University Press.

Klein, J. 2020. “Francis Bacon.” In *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*, edited by E. N. Zalta. https://plato.stanford.edu/archives/fall2020/entries/francis-bacon/.

Krakauer, J. W., A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel. 2017. “Neuroscience Needs Behavior: Correcting a Reductionist Bias.” *Neuron* 93, no. 3: 480–490.

Lee, S. Y., K. Kozalakis, F. Baftizadeh, et al. 2024. “Cell-Class-Specific Electric Field Entrainment of Neural Activity.” *Neuron* 112, no. 15: 2614–2630.e5.

List, C., and P. Menzies. 2017. “My Brain Made me Do It: The Exclusion Argument Against Free Will, and What's Wrong With It. In Beebee, Hitchcock & Price.” In *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press.

Marr D., and T. Poggio. 1976. “From Understanding Computation to Understanding Neural Circuitry.” [AI Memo 357]. MIT Artificial Intelligence Laboratory.

Meincke, A. S. 2019. “Autopoiesis, Biological Autonomy and the Process View of Life.” *European Journal for Philosophy of Science* 9: 1–16.

Menzies, P., and C. List. 2010. “The Causal Autonomy of the Special Sciences.” In *Emergence in Mind*, edited by G. Macdonald and C. Macdonald, 108–129. Oxford University Press.

Miller, P. 2016. “Dynamical Systems, Attractors, and Neural Circuits.” *F1000Research* 5: F1000 Faculty Rev-992.

Millikan, R. G. 1984. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. MIT Press.

Millikan, R. G. 1995. “Pushmi-Pullyu Representations.” *Philosophical Perspectives* 9: 185–200.

Mitchell, K. J. 2023a. *Free Agents – How Evolution Gave Us Free Will*. Princeton: Princeton University Press.

Mitchell, K. J. 2023b. “The Origins of Meaning – From Pragmatic Control Signals to Semantic Representations.” https://doi.org/10.31234/osf.io/dfkrv.

Nicholson, D. J., and J. Dupré, eds. 2018. *Everything Flows. Towards a Processual Philosophy of Biology*. Oxford University Press.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pérez-Ortega, J., A. Akrouh, and R. Yuste. 2024. “Stimulus Encoding by Specific Inactivation of Cortical Neurons.” *Nature Communications* 15, no. 1: 3192.

Pessoa, L. 2022. *The Entangled Brain: How Perception, Cognition, and Emotion Are Woven Together*. MIT Press.

Pezzulo, G., and C. Castelfranchi. 2007. “The Symbol Detachment Problem.” *Cognitive Processing* 8, no. 2: 115–131.

Piccinini, G. 2022. “Situated Neural Representations: Solving the Problems of Content Hypothesis and Theory.” *Frontiers in Neurorobotics* 16: 846979.

Pinotsis, D. A., G. Fridman, and E. K. Miller. 2023. “Cytoelectric Coupling: Electric Fields Sculpt Neural Activity and “Tune” the Brain's Infrastructure.” *Progress in Neurobiology* 226: 102465.

Pospisil, D. A., M. J. Aragon, S. Dorkenwald, et al. 2024. “The Fly Connectome Reveals a Path to the Effectome.” *Nature* 634, no. 8032: 201–209. https://doi.org/10.1101/2023.10.31.564922.

Potter, H. D., and K. J. Mitchell. 2022. “Naturalising Agent Causation.” *Entropy* 24, no. 4: 472.

Prigogine, I., and I. Stengers. 1984. *Order Out of Chaos: Man's New Dialogue With Nature*. Bantam Books.

Raja, V., and M. L. Anderson. 2021. “Behavior Considered as an Enabling Constraint.” In *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience*, edited by F. Calzavarini and M. Viola, 209–232. Springer Nature.

Randi, F., A. K. Sharma, S. Dvali, and A. M. Leifer. 2023. “Neural Signal Propagation Atlas of Caenorhabditis Elegans.” *Nature* 623, no. 7986: 406–414.

Robinson, H. 2023. “Dualism.” In *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*, edited by E. N. Zalta and U. Nodelman. https://plato.stanford.edu/archives/spr2023/entries/dualism/.

Robson, D. N., and J. M. Li. 2022. “A Dynamical Systems View of Neuroethology: Uncovering Stateful Computation in Natural Behaviors.” *Current Opinion in Neurobiology* 73: 102517.

Roli, A., J. Jaeger, and S. A. Kauffman. 2022. “How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence.” *Frontiers in Ecology and Evolution* 9: 806283.

Rosas, F. E., B. C. Geiger, A. I. Luppi, et al., 2024. “Software in the Natural World: A Computational Approach to Emergence in Complex Multi-Level Systems.” arXiv Preprint arXiv:2402.09090.

Ross, L. N. 2023. “The Explanatory Nature of Constraints: Law-Based, Mathematical, and Causal.” *Synthese* 202, no. 2: 56.

Ross, L. N., and D. S. Bassett. 2024. “Causation in Neuroscience: Keeping Mechanism Meaningful.” *Nature Reviews Neuroscience* 25, no. 2: 81–90.

Ross, L. N., V. Jirsa, and A. McIntosh. 2024. “The Possibility Space Concept in Neuroscience: Possibilities, Constraints, and Explanation.” [Preprint] accessed 2024-10-10. https://philsci-archive.pitt.edu/id/eprint/23682.

Rule, M. E., T. O'Leary, and C. D. Harvey. 2019. “Causes and Consequences of Representational Drift.” *Current Opinion in Neurobiology* 58: 141–147.

Rusakov, D. A., L. P. Savtchenko, and P. E. Latham. 2020. “Noisy Synaptic Conductance: Bug or a Feature?” *Trends in Neurosciences* 43, no. 6: 363–372.

Sanborn, A. N., J. Q. Zhu, J. Spicer, et al. 2024. “Noise in Cognition: Bug or Feature?” *Perspectives on Psychological Science* in press.

Seibt, J. 2016. “Process Philosophy.” In *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, edited by E. N. Zalta. https://plato.stanford.edu/archives/win2016/entries/process-philosophy/.

Semedo, J. D., E. Gokcen, C. K. Machens, A. Kohn, and B. M. Yu. 2020. “Statistical Methods for Dissecting Interactions Between Brain Areas.” *Current Opinion in Neurobiology* 65: 59–69.

Semedo, J. D., A. Zandvakili, C. K. Machens, M. Y. Byron, and A. Kohn. 2019. “Cortical Areas Interact Through a Communication Subspace.” *Neuron* 102, no. 1: 249–259.

Shea, N. 2018. *Representation in Cognitive Science*. Oxford University Press.

Sherrington, C. S. 1910. “Flexion-Reflex of the Limb, Crossed Extension-Reflex, and reflex stepping and standing.” *Journal of Physiology* 40: 28e121.

Shine, J. M. 2023. “Neuromodulatory Control of Complex Adaptive Dynamics in the Brain.” *Interface Focus* 13, no. 3: 20220079.

Shine, J. M., E. J. Müller, B. Munn, J. Cabral, R. J. Moran, and M. Breakspear. 2021. “Computational Models Link Cellular Mechanisms of Neuromodulation to Large-Scale Neural Dynamics.” *Nature Neuroscience* 24, no. 6: 765–776.

Siemian, J. N., M. A. Arenivar, S. Sarsfield, C. B. Borja, C. N. Russell, and Y. Aponte. 2021. “Lateral Hypothalamic LEPR Neurons Drive Appetitive but Not Consummatory Behaviors.” *Cell Reports* 36, no. 8: 109615.

Silberstein, M. 2021. “Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences 2.0.” In *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience*, edited by F. Calzavarini and M. Viola, 363–393. Springer Nature.

Silberstein, M., and A. Chemero. 2013. “Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences.” *Philosophy of Science* 80, no. 5: 958–970.

Sinnott-Armstrong, W. 2019. “Contrastive Mental causation.” *Synthese* 198, no. Suppl 3: 861–883.

Suzuki, M., C. M. Pennartz, and J. Aru. 2023. “How Deep Is the Brain? The Shallow Brain Hypothesis.” *Nature Reviews Neuroscience* 24, no. 12: 778–791.

Taylor, N. L., C. J. Whyte, B. R. Munn, et al. 2024. “Causal Evidence for Cholinergic Stabilization of Attractor Landscape Dynamics.” *Cell Reports* 43, no. 6: 114359. https://doi.org/10.1016/j.celrep.2024.114359.

Thiele, A., and M. A. Bellgrove. 2018. “Neuromodulation of Attention.” *Neuron* 97, no. 4: 769–785.

Thura, D., J. F. Cabana, A. Feghaly, and P. Cisek. 2022. “Integrated Neural Dynamics of Sensorimotor Decisions and Actions.” *PLoS Biology* 20, no. 12: e3001861.

Tinbergen, N. 1963. “On Aims and Methods of Ethology.” *Zeitschrift für Tierpsychologie* 20: 410–433.

Tse, P. U. 2013. *The Neural Basis of Free Will: Criterial Causation*. MIT Press.

Tseng, P., and T. Cheng. 2024. “Causal Prominence for Neuroscience.” *Nature Reviews Neuroscience* 25, no. 8: 591–591.

Tsimring, L. S. 2014. “Noise in Biology.” *Reports on Progress in Physics* 77, no. 2: 026601.

van Bree, S., D. Levenstein, M. Krause, B. Voytek, and R. Gao. 2024. “Decoupling Measurements and Processes: On the Epiphenomenon Debate Surrounding Brain Oscillations in Field Potentials.” https://doi.org/10.31234/osf.io/knjfw.

Winning, J., and W. Bechtel. 2018. “Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control.” *Minds and Machines* 28: 287–310.

Woodward, J. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yoshihara, M., and M. Yoshihara. 2018. “‘Necessary and Sufficient’ in Biology Is Not Necessarily Necessary – Confusions and Erroneous Conclusions Resulting From Misapplied Logic in the Field of Biology, Especially Neuroscience.” *Journal of Neurogenetics* 32, no. 2: 53–64.

Zutshi, I., A. Apostolelli, W. Yang, et al. 2024. “Hippocampal Neuronal Activity Is Aligned With Action Plans.” bioRxiv : The Preprint Server for Biology, 2024.09.05.611533. https://doi.org/10.1101/2024.09.05.611533.