Developing a Naturalistic Metaphysics for Biological Agency

By

Henry D. Potter

A thesis submitted in partial fulfilment of

the requirements for the degree of

Doctor of Philosophy

2025

The Smurfit Institute of Genetics

School of Genetics & Microbiology

Trinity College Dublin

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

This thesis contains published and unpublished work carried out in collaboration with other researchers. Their contribution is duly acknowledged at the beginning of the relevant chapters.

Henry Denis Potter

Summary

This thesis contributes to the project of naturalising agency, where agency can be broadly understood as the ability to choose and control one's actions on the basis of one's own purposes, goals, or reasons. This notion of agency is central to the basic phenomenology of our everyday existence as human beings, to how we come to understand ourselves and the world around us, and to the social, moral and legal practices that structure our societies. It is also fast becoming a critical concept for addressing emerging issues in the ethics and design of artificial intelligence.

Yet, despite its profound existential, societal and ethical importance, we continue to lack a robust understanding of the concept of agency from the perspective of the natural sciences. One of the main reasons for this is a widespread and long-standing belief that agency, and related concepts such as freedom, purpose, and meaning, carry with them certain metaphysical commitments that are fundamentally incompatible with the picture of the world as it is given to us by the natural sciences (what is sometimes known as a naturalistic worldview or metaphysics). It is thought that these agential concepts presuppose certain things about the universe that our best scientific theories in physics, biology, and neuroscience simply deny.

These naturalistic suspicions about agency have led many scientists and philosophers to conclude that our intuitive concept of agency must therefore be an *illusion*—a non-natural, unscientific notion that exists in our phenomenology, but not in the 'real world'. On this basis, some have even called for a radical overhaul of some of our most basic social, moral, and legal practices.

In this thesis, I push back on these conclusions by integrating evidence from across the natural sciences to develop a novel, empirically grounded account of biological agency. Specifically, I focus in on three core features (or *types*) of agency that are commonly claimed to be mysterious and unscientific. These are: (i) macroscopic, agent-level control, (ii) purposive or mental causation, and (iii) the ability to 'do otherwise' and make genuine choices. I approach each of these features within the context of a series of challenges for agency that have emerged out of the philosophy of action, behavioural neuroscience, the philosophy of free will, and evolutionary developmental biology.

First, I consider a common view within the philosophy of action literature that 'agent causation'—i.e., the sort of emergent, macroscopic causal power that would allow agents *themselves* to be the causal source of their own actions—is not compatible with a scientific

worldview. I draw on evidence from thermodynamics and systems biology to develop an eight-criterion framework which, I argue, collectively describes a naturalistic theoretical system that exhibits an entirely non-mysterious form of agent causation.

Second, I consider the claim that neuroscience has *shown* that our behaviours are ultimately the product of neural and biochemical activity inside of our brains and bodies, and that our mental states and subjective experiences are therefore mere epiphenomena that play no causal role in determining how we behave. I integrate philosophical literature on causation with recent empirical work in neuroscience to argue for a more expansive view of causation in the brain, which includes temporally extended, macroscopic, and semantic forms of causation. Under this causal schema, I show that behaviour is primarily the product of what is *meaningful* to these systems (and not the specific physical or biochemical details by which that meaning is transiently realised).

Third, I turn to the claim—common in the philosophical literature on free will—that biological systems like us are not free to *choose* how we will behave because either physics supports the view that our choices are pre-determined by historical factors beyond our control, or it supports the view that there is some fundamental randomness involved in how we behave, but that this simply makes our choices and actions mere matters of chance or luck. I show, first, that modern physics does *not* support the view that our choices are pre-determined by historical factors and conditions. Then, I draw on evidence from neuroscience, psychology and decision theory, to develop an entirely naturalistic model of deliberative decision-making in which it is both physically open to the agent to take more than one different action and sufficiently under their control to choose which one of those actions they will take. I argue that, on this view, an agent's choices are neither pre-determined nor lucky.

Taken together, I conclude that these arguments provide a naturalistic framework for thinking about living systems and their behaviour which resolves many of the existing scientific concerns about agency and shows how many of the concept's core features are, in fact, perfectly consistent with a modern scientific worldview. I then end by considering some recent attempts to extend a similar concept of agency into the field of evolutionary developmental biology by claiming that 'developing organisms actively and purposively control their own development'. I appeal to the wider evidence base in developmental and evolutionary biology to argue that—when interpreted literally—these claims are not empirically well-supported.

Acknowledgements

First and foremost, I am deeply grateful to my supervisor, Kevin Mitchell, for making the questionable decision to allow a philosopher into the Genetics department and for giving me the opportunity to start this PhD. Kevin's breadth of knowledge and his passion for interdisciplinary synthesis are as inspiring as they are intimidating. His approach to research will leave a lasting impression on me as a model of the kind of researcher I hope to be. Thank you, Kevin, for your constant support and guidance over these past four years. Thank you for always believing in me, even when I didn't believe in myself. And thank you for the many, many hours of stimulating discussion. I've enjoyed it tremendously.

I would also like to thank my thesis committee, Aoife McLysaght and Shane O'Mara, for agreeing to take on this slightly unusual project and for their helpful, reassuring feedback throughout the process.

It's a cliché, but no less true, that this thesis would simply not exist without the support and encouragement of friends and family. To my Dublin-based friends—especially the members of Margaret's Boys—thank you for the pints and for the laughs (in that order).

Special thanks to all the friends who made the trip over to visit me: Luke Reilly, Harry Leitch, Zahra Tengra, Nicki Betz, Michael Keeling, Amy de Wolf, Clare Wall, Becky Wheeler, Derek Spencer, Chris Brook, Ben Pegler, Kim Berisha, Celyn Morris, Sam Lister, Andrew Lavis, Sarah Fisher, Charlie Tomlin, Sophie Moss, and Josh Cullen. I can't tell you how much I appreciate those visits, and how much those moments of familiarity have meant to me. I feel very lucky to have you as mates.

To my family, thank you for always encouraging me to pursue what makes me happy, and for your repeated attempts to understand what it is I've been doing over here. To my brother Elliot, my sister Kate, and our dog Otis: thank you for your care and for always making me laugh. To my dad, Ian, thank you for teaching me to be intellectually curious and skeptical, and to love a good argument. To my mum, Tara, thank you for showing me the value of perseverance, resilience, and kindness.

Finally, Emma: thank you for being you—by far the best discovery of this PhD.

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction

A lot of what happens in the world *just happens*. Rivers flow, raindrops fall, and clocks tick. Human action, however, *feels* different. We experience ourselves not merely as objects in motion, *undergoing* change, but as *agents*, capable of causing change and being in control of what we do. We seem to make choices about when and how to act in the world. And these feel like choices that *we* come to and enact, on the basis of our own subjective goals, purposes, values, beliefs, reasons, motivations, and emotions. Unlike the river or the clock, our behaviours do not appear to be automatic and inevitable, or imposed on us by some external or foreign force—things that just *happen*. Quite the opposite, in fact: they typically feel like things that we actively *do*; we feel like the authors of our own story. Consequently, not only do we have a say in how our own lives unfold, there is also a sense in which, in some small part, we appear to have a say in how the world around us unfolds.

Aristotle framed this idea in terms of our behaviours and their effects being "up to us" (Meyer, 2014). In contemporary philosophy, it is typically referred to as having the 'power to act' (Schlosser, 2019; Glasscock & Tennenbaum, 2023). In this thesis, I will mainly use the term *agency*—which I use, broadly, to refer to the ability of an entity to choose and control its actions on the basis of its own purposes, goals or reasons (more on this later).

Such agency is clearly central to our basic phenomenology and to how we understand ourselves. Crucially, though, it is not an all-or-nothing phenomenon—as might be implied by any sort of rigid definition. Instead, we regularly discriminate between different *degrees* of control and 'up to us'-ness. Consider, for instance, the difference between an individual who steals a loaf of bread because it is the only option to feed their family and one who steals the bread simply for fun. Intuitively, both individuals exercise *some* level of agency in how they act, but there is a clear sense in which the one acting out of desperation has *less* control over the situation than their counterpart. Likewise, someone suffering from a drug addiction or a mental illness is still generally assumed to have *some* capacity to choose and control how they behave on a day-to-day basis, but we consider this capacity to increase as their health improves. Our intuitive concept of agency is therefore fundamentally a scalar or graded notion: the extent to which individuals are in control of what they do appears to us to vary along a continuum, rather than being some sort of fixed, binary property that they either have or do not have.

What these examples illustrate, too, is that the notion of agency is not merely an inconsequential feature of our phenomenology. How we conceptualise what agency *is*, where we think it applies, and what factors we take to limit it all have demonstrable real-world consequences. In most societies, a healthy person or the carefree bread-thief is far more likely to be condemned, blamed, or punished for their actions than their mentally unwell or starving counterparts, precisely *because* of the apparent difference in agency. In fact, the way that we understand and apply the concept of agency has practical implications for *many* different aspects of our lives. As we will see in **Section 1.1**, it is central to how we perceive and make sense of the world around us, to the social, moral and legal practices that structure our societies, and to many of our behavioural and cultural norms. Agency is also fast becoming a critical concept for addressing emerging issues in the ethics of artificial intelligence and animal rights.

Yet, despite its profound existential, societal and ethical importance, we continue to lack a robust understanding of the concept of agency from the perspective of the natural sciences (Froese & Taguchi, 2019; Ball, 2023a). From a biological standpoint—where organisms are typically framed in mechanistic or evolutionary terms—the idea that living systems are genuine agents, acting purposively in pursuit of intrinsic goals, remains contentious and theoretically elusive. From the perspective of physics, it is often dismissed outright as mysterious and unscientific (Greene, 2021; Hossenfelder, 2022).

Part of the reason for this is a widespread and long-standing belief that agency and related concepts such as freedom, purpose, and meaning carry with them certain metaphysical commitments that are thought to be fundamentally incompatible with the picture of the world as it is given to us by natural science (what is sometimes known as a naturalistic worldview or metaphysics). These concepts are thought to presuppose certain things about the universe that our best scientific theories appear to deny or have no need for.

These metaphysical suspicions have led many scientists and philosophers to conclude that our intuitive concept of agency must in fact be an *illusion*—a non-natural, unscientific notion that exists in our phenomenology, but not in the 'real world' (Kim, 1993; Smilansky, 2000; Wegner, 2002; Pereboom, 2005, 2014; Levy, 2011; Caruso, 2012; Harris, 2012; Greene, 2021; Hossenfelder, 2022; Sapolsky, 2023). For these skeptics, it is not just that we currently *lack* a scientifically plausible account of agency, the claim is that such an account is not *possible* within a naturalistic framework. They conclude that living systems must therefore not *really* be controlling our behaviours on the basis of our own intrinsic purposes, goals and reasons; nothing in the world must *really* be 'up to us'; we must not have any *real* say in how the events around us unfold. It just feels like we do—and there is

instrumental utility in treating one another as *though* we do (Dennett, 1987). But, in reality, any feeling of agency is ultimately illusory. Everything that happens, just 'happens'. We are, in a sense, just like the river and the clock: mere passive observers.

On this basis, some have argued that we ought to re-evaluate our basic image of ourselves, how we approach interpersonal relations, and even the way that we structure our societies (Harris, 2012; Pereboom, 2014; Caruso, 2021; Sapolsky, 2023; but cf. Smilansky, 2000). If there really is no difference between a healthy person and one suffering from a mental illness in terms of their degree of agency and control over how they act (because, for both, it is illusory), then the way in which we currently mitigate blame and punishment for some individuals, under certain circumstances, and not for others, would seem to be inappropriate.

The aim of this thesis is to push back on these conclusions by contributing to recent multidisciplinary efforts to naturalise agency (Juarrero, 1999; Di Paolo, 2005; Barandiaran et al., 2009; Steward, 2012; Tse, 2013; Walsh, 2015; Moreno & Mossio, 2015; Dennett, 2015; Ellis, 2016; Fulda, 2017; Mitchell, 2018b, 2023a; Ball, 2023a, 2023b; Froese, 2023; Jaeger, 2024). Very broadly, to naturalise a concept is to provide an account of that concept that is well aligned with the picture of the world as it is given to us by natural science (Papineau, 2023). It is an attempt to reconcile one's intuitive, 'folk', or philosophical understanding of a concept with the empirical evidence and established natural principles of modern science. In doing so, one would hope to formalise the intuitive concept into a valid and useful scientific construct and, in the process, gain a more rigorous, precise, generalisable, and possibly even operational understanding of the phenomenon itself.[1]

My overall aim in this thesis, then, is to ask: can we reconcile our intuitive concept of agency with a modern scientific understanding of the world, and of biological systems in particular?

To approach this question, I focus in on three specific features (or *types*) of agency, which I take to be central to our intuitive understanding of the phenomenon, but that are commonly claimed to be mysterious and unscientific. These are: (i) macroscopic, agent-level causation, (ii) mental or purposive causation, and (iii) the ability to 'do otherwise' and make genuine choices. In a series of independent but interrelated papers, I draw on recent empirical work from across the natural sciences to argue that each of these features of agency *is*, in fact, perfectly consistent with a modern scientific (particularly, biological)

[1] Specifically, then, this thesis is interested in an *ontological* form of naturalism that "is concerned with the contents of reality" rather than a purely *methodological* naturalism which focuses on the best way to investigate reality (Papineau, 2023).

worldview. In doing so, I hope to develop a naturalistic framework for thinking about living systems and their behaviour which can lay the groundwork for a more empirically informed understanding of agency—one which, crucially, *complements* (rather than eliminates) many of the core aspects of our intuitive and philosophical understanding of the phenomenon.

There are therefore two main motivations behind the research presented in this thesis. The first is to help develop a scientifically informed understanding of a phenomenon that has, as yet, lacked such an understanding, despite being an ever-present aspect of our lived existence. This is necessary for pushing back on certain forms of skepticism about agency and their apparently drastic existential and practical repercussions (e.g., Pereboom, 2014; Sapolsky, 2023). Moreover, an empirically grounded concept of agency should also offer new resources and provide an alternative perspective on a range of long-standing questions within the philosophy of action and free will. As will be a recurring theme in this thesis, these literatures are rife with what I will contend are outdated assumptions about what is permissible within a naturalistic metaphysics and what is not. Hence, re-assessing and dislodging some of these old assumptions, as I plan to, promises to open up new avenues of exploration within these debates.

The second motivation is to help establish agency as a useful scientific construct that can be both studied and applied, most notably in the biological sciences. As I will explore in the next section of this introduction, agency—and corollary notions such as goals, purpose, choice, and meaning—are currently, at best, considered *instrumental* concepts within mainstream contemporary biology. Such instrumentalism holds that it is epistemically useful to treat certain complex systems (e.g., organisms) *as if* they were agents, due to the exceeding complexity of their internal workings (Okasha, 2023). However, this is taken to be an explicitly pragmatic move, with no commitment to the idea that these systems really *are* agents. Within the biological sciences, this approach is characterised by the term 'teleonomy' (Pittendrigh, 1958; Mayr, 1974; Dresow & Love, 2023). In cognitive science and philosophy, it is known as the intentional stance (Dennett, 1987). One of the main objectives of this thesis is to motivate a move beyond these forms of instrumentalism. Doing so will enable a more nuanced and rich exploration of the concept of agency, which could stand to benefit not only neuroscientists, geneticists and other biologists, but also researchers in the social sciences and in the fields of artificial life and artificial intelligence. Indeed, a more empirically grounded, theoretical understanding of agency—with all of its gradations—has already been explicitly recognised as necessary for progress in areas as disparate as biological evolution (Lala et al., 2015; Walsh, 2015), organismal development

(Levin, 2022; Nadolski & Moczek, 2023; Snell-Rood & Ehlman, 2023), AI ethics (List, 2021; Andrada et al., 2023) and the problem of Artificial General Intelligence (Roli et al., 2022), public health and wellbeing (Lorimer et al., 2022), improving education (Louis & Khalifa, 2018), and even for understanding criminological risk (Piquero, 2021). Here, I hope to help lay down some of the groundwork for such an account.

The remainder of this introduction will be structured as follows. In **Section 1.1**, I reflect on the significance that a more naturalised understanding of agency could have for many areas of science, philosophy, and wider society. There are two aspects to this. First, I note how the way in which we currently think about and conceptualise agency *already* plays a hugely consequential role in how we perceive the world, in how we treat one another and our physical surroundings, and in the way in which we currently structure our legal, judicial, and political systems. I suggest that these aspects of our lives could therefore stand to gain from an empirically grounded investigation of agency that may offer new insights that help to refine and nuance the intuitive notion of agency at play here, and potentially even make it amenable to systematic experimentation. Second, I explore several areas—including artificial intelligence research, origin of life studies, and the growing field of systems biology—where a naturalised concept of agency is currently lacking but could do some useful work.

In **Section 1.2**, I then introduce the empirical challenges currently standing in the way of a fully naturalised understanding of agency. As mentioned above, in this thesis I limit my focus to three specific features (or *types*) of agency that have been the subject of scientific skepticism: (i) causal sourcehood, (ii) mental (or semantic) causation, and (iii) genuine choice (or 'the ability to do otherwise'). In this section, I explain why each of these features of agency is often considered to be metaphysically suspicious. Where relevant, I also introduce some of the recent attempts in the literature to naturalise these features of agency.

In **Section 1.3**, I summarise the discussion. I then set out the thesis's methodological approach for addressing these challenges.

In **Section 1.4**, I provide an overview of each thesis chapter, outlining each one's motivating research question and corresponding argument.

1.1. The Significance of Agency

1.1.1. The Role of Agency in Everyday Life

The concept of agency is not a mere incidental quirk of human phenomenology, on which nothing *really* hangs. Quite the opposite, in fact. How we think about and understand agency—as it applies to human behaviour in particular—has a demonstrable effect on many critical aspects of our day-to-day lives.

Even in something as basic as visual perception, for example, many philosophical and scientific traditions have independently stressed that the contents of our perceptions appear to be fundamentally shaped by some sort of pre-reflective or tacit notion of our own agency. In psychology, for instance, J. J. Gibson (1979) influentially argued that what we perceive is, in fact, a field of what he called *affordances*: the set of opportunities for action *afforded* to us by the state of the environment we encounter. On this view, organisms like us do not perceive our surroundings as impersonal shapes and objects. Instead, we perceive the world in terms of what we can *do* with it—or, at least, what we tacitly *take ourselves* to be able to do with it. Perception is therefore considered to be an intrinsically action-oriented phenomenon: for a human to perceive a chair just *is* to perceive an opportunity for sitting, in a manner that would presumably not be the case if an elephant were to perceive the same chair. According to this view, an implicit, pre-reflective notion of our own agency is therefore built right into the heart of perception itself.

Similar ideas have repeatedly appeared in both classical philosophy of mind and in more recent work in the cognitive sciences. Both Kant (1781/1929) and Hegel (1807/1977), for example, explicitly rejected the idea of perception as a form of passive reception, gesturing instead toward a view in which the world 'shows up for us' in a way that is guided by or disclosed to us through our own practical capacities. Twentieth-century phenomenologists made the point explicit: most notably, Merleau-Ponty's *Phenomenology of Perception* (1945/1970) portrays vision as "I can" rather than "I think" (Toadvine, 2023). In analytic circles, Ryle (1949) and Wittgenstein (1953) emphasised action-readiness in recognising objects ("seeing-as"), while Gibson's ecological legacy and focus on *affordances* was taken up and developed further by Reed (1996) and Chemero (2009). More contemporary enactivists (e.g., Varela, Thompson & Rosch, 1991; O'Regan & Noë, 2001; Di Paolo, 2005) similarly argue that perceptual content is *constituted by* the sensorimotor contingencies an embodied agent can enact in that moment (i.e., 'if I do this, then that will change'). Across all of these traditions, however, the recurring theme

is the same: to perceive is to view the world in light of one's own possible *doings* (that is, to see it through the lens of some implicit notion of agency).

These insights are being increasingly corroborated by recent work in cognitive neuroscience. The increasingly influential theories of predictive processing and active inference (Friston, 2005, 2010; Clark, 2013, 2015), for example, argue that what visual perception *is* is as an active, top-down process in which the brain continuously generates predictions about the causes of its sensory inputs and then updates those predictions in light of new evidence. Crucially, these predictions are said to be shaped not only by prior knowledge but also by the perceiver's internal *self*-model (Friston, 2018)—which includes an internal model of their own capacities for action (Nave et al., 2022). In predicting how the world will (or should) appear, the idea is that the brain is already inferring and implicitly modelling what the body might be able to *do* within it. Thus, on these views, perception is again assumed to be inextricably tied to one's implicit concept of their own agency: we perceive the world not as it is, but in terms of how we (tacitly) view ourselves to be able to act within it.

How we conceptualise and understand our own agency therefore quite literally shapes how we see the world around us—what it *affords* to us. This, in turn, impacts how we behave. A similar effect is evident in social psychology, where researchers have found that how an individual views and consciously reflects on their own agency—what is sometimes referred to as the subject's 'perceived self-efficacy' (Bandura, 1977)— can have observable effects on which actions they choose to pursue (Bandura, 1977), how much cognitive effort they put into those actions (Bandura, 1982), and how much they persevere with the action in the face of obstacles and other difficulties (Weinberg et al., 1979). Higher perceived self-efficacy has been found to significantly correlate with greater academic perseverance (Multon et al., 1991), improved work-related performance (Stajkovic & Luthans, 1998), and better performance in sport (Mortiz et al., 2020).

Moreover, it is not only how we conceptualise our *own* agency that influences our behaviour. How we attribute agency to *others* can also shape and inform how we behave. In part, this is because the concept of agency seems to play a central role in how we come to parse and make sense of the world around us. A wealth of evidence from developmental psychology, for example, suggests that even very young infants deploy some minimal notion of agency in the way that they learn about the world. Several studies have suggested that infants naturally carve up their surroundings into objects (perceived to be) capable of moving under their own power and objects that need to be pushed, pulled, or otherwise interacted with in order to move (Premack, 1990; Spelke et al., 1996). Infants as

young as twelve months even seem to sometimes attribute intrinsic goals to these self-moving objects, interpreting them as a sort of purposive agent acting in pursuit of a goal, as a way of understanding and predicting their behaviour (for overviews, see Gergely & Csibra, 2003; Steward, 2009). In one well-known example, infants were experimentally habituated on an animated scene in which a small circle 'jumps' over a rectangular barrier to reach another object (Gergely et al., 1995). The authors found that, when the barrier was then removed in the test conditions, infants displayed longer looking time in trials where the circle continued to 'jump' unnecessarily (i.e., despite there being no barrier to 'jump' over), as compared with ones where the circle moved directly toward the other object. Given that longer looking times typically indicate surprise or a violation of expectation, the authors suggested that the best explanation for these findings was that the infants were expecting the circle to act efficiently based on a perceived 'goal' to reach the other object. They therefore concluded that the infants must be deploying a concept of agency to make sense of the observed behaviour—interpreting the circle's movements as, in some important sense, *purposeful* and *goal-directed*.

Given that we appear to cognitively carve the world up in this way—into objects and entities with agency and those without—it seems natural to expect that our behaviours will be directly influenced by the concept of agency we are implicitly deploying (and where we are applying it). That is because we make demonstrably different behavioural choices in the presence of (what we perceive to be) genuine agency than we do otherwise. Consider, for example, the common experience of, say, mistaking a leaf blowing in the wind for a mouse, or a strand of hair tickling the back of our neck for a spider. In these situations, we typically find ourselves acting (or planning to act) in dramatically different ways *before* we realise our mistake than we eventually do *afterward*.

Human history is replete with far more consequential examples in which these shifts in how we attribute agency to the world prompted profound changes to our behavioural and cultural practices. Natural phenomena such as storms, droughts, and plagues, for example, were historically interpreted as the actions of hidden agents (gods, spirits, or other supernatural forces), acting in accordance with their own goals, purposes, and reasons (see Riskin, 2016). Entire behavioural rituals developed around these (mis)attributions of agency, including practices like rain dances, appeasement ceremonies, and even sacrificial offerings—which have since been abandoned within most cultures. Similarly, our ancestors often considered what we would now classify as mental illness to be cases of demonic possession or some sort of spiritual corruption (see Sapolsky, 2023). Again, these

(mis)attributions of agency shaped how these individuals were treated, often leading to morally reprehensible practices such as exorcisms and social alienation.

Furthermore, it is not just *where* we perceive agency to exist in the world that is important. What *type* or *degree* of agency we attribute to one another also has profound consequences for how we behave. In legal and interpersonal contexts, for example, we now mitigate blame and punishment for children and people suffering with mental illness on the assumption that they lack the epistemic and agentic capacities required to fully recognise and act in accordance with moral and social norms. Historically, too, slaveowners and broader society often sought to "justify" the horrors of the institution of slavery and the lack of moral patiency afforded to enslaved people, by selectively (mis)attributing a lack of certain types of epistemic or moral agency to these people (Pleasants, 2010).

How we think about and conceptualise agency therefore clearly impacts how we behave and how we treat one another, sometimes in profoundly important ways. In fact, it would not be an exaggeration to say that almost every aspect of human society is built on an intuitive assumption that we *are* agents of a certain sort. As Aristotle (1985) observed in his *Nicomachean Ethics*, the very ideas of morality and virtue themselves—which ultimately underlie pretty much all of our social practices of holding each other responsible for the things that we do—necessarily presuppose a notion of agency, that is, of there being a distinction between voluntary acts and involuntary movements, or between outcomes and events that are 'up to' an agent and those that are not. It would be conceptually incoherent, for instance, to truly judge someone for how they acted—to praise, blame, reward or punish them—if you did not consider them to be somehow in control of that behaviour. Or, indeed, to feel pride or shame for one's *own* behaviour if behaving in that way was never something that was 'up to you' in the first place. Moreover, it seems conceptually incoherent to judge a person for their virtues and vices—to admire, resent, respect, hate, or even love them—if they had no real say in their being the type of person they are. We do not admire the sun for feeding our crops or blame the rain for causing floods because we (no longer) think they are exercising any sort of agency in bringing about these outcomes; they are just things that happen.

Our legal systems are also fundamentally built on a basic conception of agency. First, in the sense that legal frameworks generally operate on the default assumption that individuals *are* agents, making reasoned choices and acting in a way they "need not have done" (Steward, 2012, p.164). Under typical circumstances, our legal system holds people accountable on the basic assumption that it *was* genuinely 'up to them' what they did; they

were in control of the action, they were not compelled to do it, and they were acting on the basis of their own reasons and intent. This is the basis for legal doctrines like the 'freedom of contract' (the ability to choose for oneself what contractual obligations to enter into and be bound by), the idea of intellectual property rights (a recognition that a person's work is something they actively did) and it is the rationale behind the dominant retributivist model of criminal justice in many Western societies (which prioritises punishment as a form of justice) (see Morse 1994).

Second, as discussed above, the legal system also goes to great lengths to discriminate the types and degree of agency one exercises in acting, and often mitigates or even excuses a person's liability in light of this. At a most basic level, we see this in the distinction between *mens rea* ("guilty mind") and *actus rea* ("guilty act")—two of the most fundamental principles of the Western legal system. Mere bodily movement (*actus rea*) is generally not considered enough to determine legal responsibility; the agent's mental state, such as their intent or state of knowledge, also matters. In criminal law, for example, defendants can put forth 'insanity' or 'diminished capacity' defences on the basis that they lack sufficient mental capacity to act with reasonable foresight or intention. There are also 'duress' and 'loss of partial control' defences, which argue that circumstances compelled the defendant to act in ways they would not otherwise have done.

Outside of the law, too, democratic political systems, consumer choice in market economies, the principle of informed consent in healthcare, and the business practices of promotions, firings, blame and reward are all core elements of society that intimately rely on the idea that humans *are* agents, making autonomous choices and controlling their actions on the basis of their own intrinsic goals and beliefs.

In sum, the point is that how we think about and understand agency is inextricably tied to, and hence hugely consequential for, our everyday existence. The payoff for developing a more naturalistic account of agency therefore extends far beyond simply validating the phenomenology of our lived experience (although this would still be a worthwhile endeavour in and of itself). How we think about the nature of our own agency is *already* playing a demonstrably important role in how we live our lives (in a way that is nuanced and graded, starting from the basic assumption that we *are* agents and then taking into consideration the ways this can vary under different circumstances). As such, developing a more rigorous understanding of the phenomenon, by investigating its biological underpinnings and establishing a functioning *science* of agency, stands to have some far-reaching implications. In particular, it should give us new insights into what *kind* of agents

we are, which can then be used to inform policies and practices that contribute to the betterment of society.

1.1.2. Agency and AI

Agency is also fast becoming a pivotal concept in the study and development of artificial intelligence (AI). As I explore in this subsection, being able to discern whether AI systems really *are* agents—whether what they do is truly 'up to them'—is soon going to be of critical importance for a range of ethical and engineering considerations that have already started to emerge and will become increasingly pressing as AI development advances.

However, our ability to detect such artificial agency is greatly complicated by the fact that these AI systems are trained at scale on human-generated data and are optimised to produce human-like outputs. What this means—as has been evident during the recent rise of Large Language Models (LLMs)—is that our intuitive instincts are no longer a reliable tool for detecting genuine agency, intelligence, and consciousness in the world. Intuition is, almost by definition, bound to fail us when it comes to assessing the agency of these highly sophisticated AI systems; behavioural analysis alone cannot serve as a reliable test for artificial agency.

Instead, the rise of AI forces us to define and operationalise the concept of agency in ways that can then support a more systematic and principled investigation of artificial agency. This will be necessary to make progress on a range of increasingly important questions in the field of AI ethics, which bear directly on issues related to legal accountability, moral assessment, and design choices. In the rest of this subsection, I will introduce two areas of AI ethics that make this need particularly vivid.

First, the issue of artificial agency throws up several questions surrounding basic accountability and responsibility practices. Historically, when machines or artefacts caused harm, responsibility could be relatively easily attributed to the human agents who either operated the machine directly or could reasonably foresee and prevent its harmful effects. Even when machines ran autonomously, their behaviour was sufficiently well-specified and predictable that accountability could be attributed to the designer or operator who failed to anticipate and/or mitigate the risks, because, in effect, these individuals had sufficient control over the machine to be justifiably held responsible for its effects (Vallor & Vierkant, 2024).

The opacity of modern AI has undermined this sort of predictability and control. Often, neither designers nor operators have it within their control to predict (and thus prevent) the potentially harmful behaviour of AI systems. This has created so-called "responsibility

gaps", wherein it is not clear whether there is moral justification for holding the designer or operator of the AI system responsible for harms they could not predict or forestall (Matthias, 2004; Vallor & Vierkant, 2024).[2]

One popular solution to these responsibility gaps is to argue that we *can* still hold the human designer/operator of these systems responsible for their effects, on the grounds that they knowingly deployed a *predictably unpredictable* machine to perform that task (Himmelreich, 2019) in a way that they "need not have done" (Steward, 2012, p.164). However, if these systems really *are* autonomous agents in their own right—making choices and taking actions in ways that are truly 'up to them'—then this justification for holding the designer/operator responsible seems harder to maintain, since responsibility tracks control and agents (by definition) control their own actions. At minimum, then, the existence of artificial agency is going to leave us with responsibility gaps once again (which we may wish to ward off by developing and using a more formalised understanding of agency to *avoid* creating AI systems with such agency in the first place). In addition, though, it may also force us to grapple with the question of whether to hold the AI system *itself* morally responsible (and possibly legally accountable) for the harm, on the basis that it qualifies as a full (moral) agent in its own right (Dattathrani & De', 2022). A more empirically grounded framework for thinking about agency is therefore going to be crucial for making these sorts of judgements.

A second, and related, area of AI ethics where questions about artificial agency are going to really matter concerns issues surrounding moral patiency: when, if ever, will AI systems merit moral *consideration* in their own right? At what point do we have a moral duty toward them? One common way of assessing the moral patiency of a system (e.g., an animal), which can be traced back to the writings of Jeremy Bentham, is to ask whether the system is capable of suffering (Danaher, 2017; Tavani, 2018). Typically, the ethical analysis of such suffering is tied to phenomenal consciousness and to feelings of pain or emotional distress (Tavani, 2018). However, one might reasonably argue that being prevented from exercising one's capacity for agency is *also* a form of suffering in and of itself. That is partially why imprisonment is considered a punishment and free-range chicken farming practices are generally seen as more ethical than factory farming. Again, then, having the theoretical and empirical resources to distinguish 'true' artificial agency from

[2] Some have even started exploiting these responsibility gaps to engage in practices of 'agency laundering', which is said to involve "obfuscating one's moral responsibility by enlisting a technology or process to take some action and letting it forestall others from demanding an account for bad outcomes that result" (Rubel et al., 2019, p.1018).

sophisticated impressions of it is going to be crucial for guiding policy decisions—and maybe even design guidelines—when it comes to these ethically high-stakes scenarios.

In sum, dealing with the rise of artificial intelligence from both an ethical and design standpoint is necessarily going to bring the concept of agency to the forefront of public and academic discourse. Not only do AI systems promise to bring a range of new agency-related moral and legal considerations to the table, but their very design also means that our intuitive tools for detecting and assessing artificial agency in the ways that will be essential for answering these questions are not suited to the task. The emergence of AI therefore demands a more naturalised—and thus operationalisable—understanding of our own agency in order to safely navigate the choppy ethical waters that lie ahead.

1.1.3. The Importance of Agency for the Life Sciences

Developing a philosophically robust, scientifically grounded concept of agency should also be of considerable value to the natural sciences themselves. Biology, for example, has had a long and complicated historical relationship with agency and related notions such as purpose, meaning, and teleology (Riskin, 2016; Ball, 2023b). When originally coining the term *biologie* in 1802, Jean-Baptiste Lamarck had initially hoped to inaugurate a distinctive field of study into the "*pouvoir de la vie*" (force of life): the capacity for intrinsic purposeful activity which he took to be the fundamental difference between living and non-living matter (Riskin, 2016, p.199). Schrödinger, too, took agency to be an essential characteristic of biological systems, as noted in his book *What Is Life?*:

> "What is the characteristic feature of life? When is a piece of matter said to be alive? When it goes on '*doing something*', moving, exchanging material with its environment, and so forth" (1944/2012, p.69, my emphasis)

Contemporary thinkers continue to argue that agency is one of *the* defining characteristics of life itself (Levin & Dennett, 2020; Ball, 2023a, 2023b; Rosslenbroich et al., 2024) and some have even suggested that it may provide a more tractable route into questions about the origins of life (Mitchell, 2023a). Yet in their day-to-day work contemporary biologists remain highly wary of using concepts like agency and purpose to describe the activities of organisms. The active, purposeful striving that, by all appearances, seems to define living systems is discussed only ever under the pretence of metaphor; rarely with any ontological weight (Riskin, 2016). This is despite the fact that it is almost impossible *not* to talk this way when describing organisms. As the somewhat dated quip, attributed to J.B.S. Haldane, goes: "teleology", i.e., the act of treating organisms as goal-directed agents, "is like a

mistress to a biologist: he cannot live without her but he's unwilling to be seen with her in public" (Mayr, 1988, p.63).

As a solution, Colin Pittendrigh (1958) proposed the term 'teleonomy' as a way for biologists to describe the *apparent* purposiveness of organisms, while avoiding the conceptual tension surrounding agency itself. In a letter to Ernst Mayr (published in Mayr, 1974), Pittendrigh wrote:

> "I wanted a word that would allow me (all of us biologists) to describe, stress or simply to allude to—without offense—this end-directedness of a perfectly respectable mechanistic system. Teleology would not do, carrying with it that implication that the end is causally effective in the current operation of the machine."

A naturalised concept of agency, which places the concept on firm scientific footing, could therefore ease this very live tension at the heart of biology and motivate a move beyond mere instrumentalism about purposiveness and goal-directedness. Doing so may offer new insights into fundamental biological questions, such as the origins of life. Some have even argued that a recognition of natural agency is necessary to support an emerging paradigm shift in evolutionary theory and developmental biology (Walsh, 2015; Lala et al., 2015; Sultan et al., 2022)—a claim I consider in detail in **Chapter 6**.

There is good reason to think that now is precisely the right time to be developing a biological concept of agency. A multitude of recent technologies and techniques have dramatically increased the scale and detail at which neuroscientists in particular can now image the brains of active, behaving animals, allowing for tens of thousands of neurons to be tracked at any one time. These advances have opened the door to large-scale, systems-level analysis as a practical possibility in the biological sciences. When researchers only had the tools to perform more component-focused, reductive experiments, there was no need for system-level concepts (such as agency) because we had no system-level data or the means to test system-level hypotheses. In light of these new technologies, however, new conceptual toolkits are needed. It would be a mistake to expect our existing concepts—built to fit the sorts of component-focused (reductionistic) data and tools we had access to at the time—to map onto this new type of data. Instead, we should look to develop new concepts and ways of thinking, of which agency is surely an important candidate, in order to capitalise on the benefits of the new technology and to ensure conceptual rigour is maintained in empirical research interpretation as we enter this new era of systems science in biology and neuroscience.

1.2. The Problem(s) of Natural Agency

So, what is the problem? Few things seem more obviously true to us than the idea that we—perhaps along with much of the living world—are agents whose behaviours are genuinely 'up to us'. We routinely seem to cause effects in the world, on the basis of the choices we make and the intrinsic motivational states (goals, beliefs, reasons) that inform those choices. There is therefore good *a priori* reason to expect that agency is a natural property of living systems—an expectation that even fuelled the original conception for the field of *biologie* itself. Given the concept's intuitive force, explanatory potential, and societal influence, one might reasonably wonder why there is not already an established science of agency. Why do biologists still prefer to keep their teleology hidden in the shadows?

A significant part of the answer, as discussed earlier, is that many believe the concept of agency to be fundamentally incompatible with our current best scientific theories. Skeptics contend that agency—and related concepts like goal-directedness, purpose, and choice—are metaphysically committed to a picture of the universe that cannot be squared with the empirical findings of physics, biology, and neuroscience. Such agency skepticism is therefore typically presented as the scientifically informed perspective, although individuals tend to arrive at its conclusions via vastly different routes. Skeptics drawing primarily from physics, for example, will typically argue that everything we do is simply the "unfolding of the given" (Smilansky, 2000, p.284): the laws of physics dictate everything that will happen in our universe and human action is no exception to this rule. Thus, we are not *really* in control of what we do (Honderich, 2002; Greene, 2021; Hossenfelder, 2022). Others who rely more on biology and evolutionary theory generally contend that it is the *purposiveness* aspect of agency that is illusory, arguing that behaviour is better explained through natural selection and genetic inheritance rather than any sort of intrinsically goal-directed agency (Cashmore, 2010; Sapolsky, 2023). Skeptics from the more neuroscientific perspective, meanwhile, often focus on the idea of conscious agency, citing empirical findings from cognitive science and neuroimaging studies that appear to imply that our conscious awareness is more like a *byproduct* of our neural processes than a locus of genuine causal power (Libet et al., 1982; Crick, 1994; Wegner, 2002; Harris, 2012).

This makes things very difficult for those engaged in the project of naturalising agency. Not only are there many different lines of scientific attack to contend with at once, these attacks also often seem to be targeting different features or aspects of agency. In this sense, the notion of agency is something of a moving target for naturalistic critiques—a 'whack-a-

mole' concept wherein, as soon as you begin to make headway on one objection, a new one pops up in its stead.

As such, in order to make the problem tractable, in this thesis I narrow my focus to just three main features (or *types*) of agency that have traditionally been the target of naturalistic critique. These are:

> i) **Causal Sourcehood** – the capacity to be the genuine causal source or locus of control for one's own choices and actions (and the effects these produce in the world).
>
> ii) **Semantic Causation** – the ability to act purposively or for reasons, in such a way that it is the content or *meaning* of one's motivational states (goals, beliefs, desires, intentions) that is making a causal difference to how they behave.
>
> iii) **Genuine Choice** – the ability to act in a way one "need not have done" (Steward, 2012, p.164), by actively choosing how to behave from among a set of physically possible courses of action.

Each of these aspects (or types) of agency carries with it certain metaphysical commitments—conditions on what the world would need to be like for that type of agency to exist—which have been argued to be incompatible with scientific naturalism. These are summarised in **Table 1**.

Feature of Agency	**Metaphysical Commitment**	**Challenge from Naturalism**
Causal Sourcehood	Emergent, macroscopic, agent-level causation.	Causal Reductionism
Semantic Causation	The causal efficacy of intrinsic meaning.	Eliminative Materialism
Genuine Choice	i) A physically open future. ii) Control over which future occurs.	i) Physical Determinism ii) Challenge from Luck

Table 1. Summary of the Problem(s) of Natural Agency: The three features of agency that will be the focus of the present thesis, alongside their corresponding metaphysical commitments and the naturalistic arguments that target those commitments.

In the rest of this section, I give a detailed overview of each of these lines of attack. For each, I introduce the feature of agency itself and explain why it is typically considered to be a core aspect of our intuitive or philosophical concept of agency. I then present the naturalistic challenge that is often taken to threaten the scientific credibility of that aspect of agency. I then briefly hint at the way in which I aim to overcome that challenge within this thesis.

1.2.1. Causal Sourcehood

One of the most basic features (or types) of agency is sourcehood: for an action to be genuinely 'up to' an agent, it is typically held that "she must be the source of her action in an appropriate way" (Caruso, 2014, p.182, see also Clarke 2003; Tognazzini, 2011; Timpe, 2016; Sartorio, 2016). Intuitively, this seems almost trivially true. But can we say more about what such sourcehood consists in?

At a minimum, sourcehood entails a specific kind of causal relationship between the agent and the action (or its effect). An event that is entirely caused by processes with no physical connection to an agent, *A*—such as a rock being dislodged during a landslide on Mars—is simply not a plausible candidate for being 'up to' *A*. Instead, foundational to the concept of agency seems to be "the idea of a causal power, a power to cause things" (Alvarez, 2013, p.103).[3] As Albert Bandura (2006) puts it, "in acting as an agent, an individual makes causal contributions to the course of events" (p.165).[4]

Still, most accounts of sourcehood demand more than mere *participation* in a causal sequence. To be the source of an action (or its effect) is to be more than just a link in the causal chain that produces it. As compatibilists about free will (e.g., Frankfurt, 1971; Fischer & Ravizza, 1998; Fischer, 2007; Dennett, 2015; Sartorio, 2016) regularly emphasise, *how* an agent 'causes things' also matters for agency. Thomas Hobbes (1651, Ch. XXI), for example, argued that we must be "free of Opposition" if our actions are to be properly considered 'up to us'. Most concretely, this means being free of external

[3] It is an open question whether omissions ought to be considered exercises of agency (see Clarke, 2014). However, it is widely accepted that, if they are, then they still require the agent to have *some* 'power to cause things' which they did not exercise.

[4] Under certain technical conceptions of agency used in anthropology (e.g., Latour, 1996), science and technology studies (e.g., Pickering, 2023), and AI research (see Johnson & Verdicchio, 2019), 'causal contributions' of this sort are both necessary *and sufficient* to ascribe agency to a system. Schlosser (2019) has labelled this the "very broad sense" of agency, wherein: "Whenever entities enter into causal relationships, they can be said to act on each other and interact with each other, bringing about changes in each other. In this very broad sense, it is possible to identify agents and agency, and patients and patiency, virtually everywhere." In this thesis, I am interested in 'narrower' usages of the term that are more common in both folk and philosophical discourse.

impediments and physical compulsion. Consider, for instance, a person bumping into you after being blown over by the wind. Or one of the subjects in Wilder Penfield's infamous brain stimulation studies who experiences their arm rising involuntarily when Penfield stimulates their motor cortex (see Mitchell, 2023a). In such cases, bodily movement is being caused by—or 'from within'—the agent, but the agent is certainly not its causal source. Instead, it is generally held that "[f]or an action to be genuinely ours, we must have causal *control* over it" (List, 2019, p.25, *my emphasis*).

Causal control is clearly a stronger condition on sourcehood than mere agent-involving or agent-internal causality, but we are still left wondering what exactly is required for an agent to exercise causal control over a choice or action (and, hence, to be its causal source). Here, we can look to the philosophy of action and the philosophy of free will literature. Since Donald Davidson's (1963) seminal paper *Actions, Reasons, and Causes*, the dominant approach in the philosophy of action has been to define the sort of causal control required for the sourcehood aspect of agency in terms of a particular *type* of causal relationship. Specifically, on this view, something is an action over which the agent has control if it is caused, in the right way, by the psychological or mental states of the agent that rationalise it.[5] Typically, these mental states get couched in terms of belief-desire pairs. Thus, an agent is said to be the source of a behaviour only when it is caused, in some way, by a combination of their desire (or goal) to bring about a particular state of affairs and a belief about how one must act in order to do that. So, for example, my drinking water is considered an action that is 'up to me' because it is caused by my desire to quench my thirst and the belief that picking up and drinking from the water bottle will achieve that.

This approach to operationalising sourcehood has been further developed in the philosophy of free will literature, largely in response to worries about certain forms of Hobbesian "Opposition" that seem to be more *internal* to the agent. Commonly cited examples include things like addictions, impulses, and phobias, all of which may plausibly satisfy the criteria for being 'actions that are caused, in some way, by the mental states of the agent', but which, intuitively, do not seem to be cases where the agent is properly in control of what they are doing.

The proposed solution—pursued vigorously by compatibilist philosophers of free will—has been to argue that the *structure* of the agent's psychology or mental states is what matters. For an agent to be the causal source of their behaviour, on these views, is for the

[5] It is important to note that Davidson's proposal came within the context of the Wittgensteinian orthodoxy of the time, which viewed reasons and causes as entirely distinct modes of explanation for understanding behaviour (Malpas, 2024). Davidson did more than anyone to bring into view the idea that reasons might be explanatorily potent precisely *because* they are causally potent.

behaviour to be caused by a particular *type* of psychological structure or mechanism. Various versions of this compatibilist account of sourcehood have been proposed. Harry Frankfurt (1971), for example, proposed an influential "hierarchical" or "mesh" account which holds that an agent is the true source of an action only when the so-called 'first-order' desire that brings about the action is *aligned* in some way with the agent's second-order desires (desires about these first-order desires). Frankfurt argues that this internal alignment of motives ensures that the agent is properly 'identified with' the causes of the action (Timpe, 2016) and thus that their source lies in the agent's own psychology rather than some sort of internal compulsion. By contrast, John Martin Fischer's (2007; also Fischer & Ravizza, 1998) compatibilist account emphasises the "reasons-responsiveness" of the causal mechanism that produces the behaviour. On this view, a behaviour is the agent's own only if the mechanism that produces it is suitably "sensitive to reasons such that, if different reasons were to bear upon it, it would respond differently, and the agent whose mechanism it is would act differently than she does act" (McKenna & Coates, 2024). This guarantees that the causal source of the action lies in the *rational* capacities of the agent, thereby ruling out phobic or addictive behaviour as agential.

However, despite their internal differences, these compatibilist attempts to operationalise causal sourcehood generally all face the same well-known problem—the so-called *Problem of the Disappearing Agent*—a challenge to which I am broadly sympathetic.[6] The *Problem of the Disappearing Agent* notes that however one conceptualises the nature of the causal mechanism or structure that produce these actions, under these views, the locus of causal agency is still always going to reside entirely *inside* the agent, i.e., with some of its parts. That is, these views are still committed to what is known as an "event-causal" model of action causation, wherein the relevant causal power always ultimately rests with a specific (set of) mental states or their neural realisers, and not with the agent *herself*. Here are some examples of philosophers forcefully raising this objection:

> "In [standard accounts of causal sourcehood], reasons cause an intention, and an intention causes bodily movements, but nobody – that is, no person – *does* anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part" (Velleman, 1992, p.461)
>
> "If the acts of an agent are caused by his wants and beliefs, how can *he*, the agent, be considered their cause?" (Goldman 1970, p.81).

[6] One exception is Ned Marksoian (1999; 2012) who proposes an agent-causal version of compatibilism that evades the 'disappearing agent' problem.

> "Intuitively, we think of agents as *carrying out* their intentions or *acting in accord with* their practical reasons, and this seems different from (simply) being caused to behave by those intentions or reasons" (Bishop, 1989, p.72, *original emphasis*)

This *Problem of the Disappearing Agent* has therefore precipitated a split in how philosophers conceptualise what it means to be 'the causal source of one's action' or to 'have causal control over it' (see Tognazzini, 2011). Event-causalists proceed with the Davidsonian picture sketched above (e.g., Kane, 1996; Fischer & Ravizza, 1998; Ekstrom, 1999, 2019; Balaguer, 2010; Franklin, 2014, 2018; Dennett, 2015; Sartorio, 2016), while so-called agent causationists adopt a view of sourcehood in which agents, *as a whole*, must somehow be capable of causing effects, *in their own right,* if they are to exhibit the sort of causal control demanded by our intuitive concept of agency (e.g., Chisholm, 1976; Reid, 1983; Markosian, 1999, 2012; O'Connor, 2000, 2009; Steward, 2012; see also Clarke, 2003).

On this agent-causal view, being an agent requires being the seat of some sort of macroscopic causal power that is not entirely reducible to, or derived from, the causal powers of one's component parts. The agent causationist Timothy O'Connor (2009), for example, describes it as an "ontologically primitive causal power" that is "at once causally dependent on microphysically-based structural states and yet ontologically primitive" (p.195).

As mentioned, I find the *Problem of the Disappearing Agent* to be compelling. In this thesis, I therefore consider agent-level causation to be conceptually necessary for the sort of causal sourcehood entailed by our intuitive notion of agency.[7] I will assume that successfully developing a naturalistic account of agency requires a proper scientific understanding of agent causation.

This assumption quickly runs into scientific concerns, however. The notion of agent causation is widely regarded as naturalistically implausible and maybe even metaphysically extravagant. Pereboom (2005, p.228), for example, states that agent causation "is not credible given our best physical theories", while van Inwagen declares that "agent-causation is a mystery" (2014, p.281).

[7] I should note that some have historically taken agent causation to refer to some sort of dualist, supernatural or magical force; a "prime mover unmoved" or "uncaused cause". For my purposes here, I just take it to mean agent-level causation: an emergent causal power that irreducibly inheres at the level of whole system. This will then be supplemented with elements of indeterminism and purposiveness in later sections.

In part, these naturalistic concerns stem from the fact that agent causation, as a concept, is wholeheartedly committed to a metaphysics of causal *emergence*—of irreducibly macroscopic modes of causation—which many consider to be completely at odds with the sort of causal reductionism that is apparently the cornerstone of our scientific worldview (Kim, 1993; Bickle, 2015; Barwich, 2021; Carroll, 2021). Let's label this the 'challenge from reductionism'.

There are two versions of the 'challenge from reductionism'. The first comes from the idea that, in a naturalistic metaphysics, relatively 'higher level' or more macroscale phenomena are always going to be ultimately derivable from—or are perhaps just a convenient level of description *for*—events and processes that are *actually* unfolding at the 'lower' or more microscopic scales of reality (e.g., Kim, 1993). Christian List (2019) captures this perspective nicely when he describes how "[i]t is widely accepted that human cognition and behaviour are ultimately the result of complex biological and physical processes in the brain and body" (p.33). The reason for this, he goes on to say, is that the worldview endorsed by most scientists and philosophers is one in which:

> "chemical processes ultimately stem from physical processes; the laws of quantum mechanics underpin the way molecules are composed of atoms and the way they interact with one another. Biological processes stem from chemical and physical processes. Think of processes such as photosynthesis or the biochemistry of cells. Biology is a product of chemistry, which, in turn, is a product of physics. Finally, psychological processes stem from physical, chemical, and biological ones" (*ibid*)

The upshot of this supposedly naturalistic worldview is that all of the 'real' physical causation in the universe "comes from the bottom up" (Mitchell, 2023a, p.152): all causal power is exhausted in the interactions that play out at the smallest scales of physical reality (whether that be quarks, strings or quantum fields). In philosophy, this idea is sometimes known as the 'causal closure of the micro-physical' (Chalmers, 2003; Kim, 2005) and it entails a form of causal reductionism that often underpins scientific and philosophical hostilities toward the notion of emergent, agent-level causation. That is because, if one accepts the causal closure of the microphysical, then macroscopic causal powers—such as those invoked by agent causationists—*must* be either reducible to microphysical interactions or simply dismissed as metaphysically incoherent (Kim, 1993, 1998, 2005; Papineau, 2002). Under such a worldview, the idea of agent-level causation simply "makes no sense" (Searle, 2001, p.82).

The second 'challenge from reductionism' for agent causation and causal sourcehood looks beyond the apparent scientific implausibility of emergent, macroscopic causation. It,

instead, emphasises the fact that as we learn more about *how* the electrochemical processes in our brains and bodies give rise to cognition and behaviour, the more the empirical evidence seems to support the very 'event-causal' picture that agent causationists explicitly reject and deem incapable of securing the causal sourcehood required for agency.

This challenge is supported by the fact that scientists can now predict and even exert control over an organism's behaviours, simply by identifying and activating specific biochemical pathways or neural circuits *within* the agent (e.g., Siemian et al., 2021; Filipowicz et al., 2022). The apparent success of this approach—of using functional decomposition and anatomical localisation to identify the causal mechanisms that appear, for all intents and purposes, to be *controlling* the organism's behaviour—seems to provide naturalistic support for an event-causal picture which (to an agent causationist) merely reduces the organism itself to a "arena" within which complicated happenings take place; it depicts living beings as simply being 'pushed around' by the chains of cause-and-effect going on *within* them. Hence, the suggestion is that, even if emergent, irreducible forms of causation *were* naturalistically plausible, our best scientific theories of behaviour would still not support the view that biological systems like us exhibit the sort of holistic, agent-level form of emergent causal control over the way that we act which many (including myself) take to be necessary for the sourcehood aspect of agency. Instead, as Stephen Hawking and Leonard Mlodinow (2010) put it, if this really is what science tells us then "it seems that we are no more than biological machines and that [agency and] free will is an illusion" (p.32).

In this thesis, I defend the idea that agent-level causal sourcehood is not only conceptually necessary for agency, but also naturalistically viable. I argue against both forms of causal reductionism discussed above, at least insofar as they are taken to apply to biological systems. Instead, I develop a view in which modern science can be seen to support both (i) the existence of emergent macroscopic causation, and (ii) the idea that such causation inheres specifically at the level of the whole organism, i.e., the agent.

1.2.2. Semantic Causation

In emphasising the importance of action 'being caused in the right way *by the agent's mental states*', Davidson and the other event-causal theories of sourcehood point us toward another core feature of the concept of agency—what I will call 'semantic causation'. I use 'semantic causation' to refer to the idea that agency is not solely about being the causal source of an effect; there is also generally considered to be some sort of subjective, purposive, intentional, rational or even mentalistic dimension to it. In other words, part of

what seems to distinguish agential from 'merely biological' or 'purely mechanistic' emergent causation is the idea that the subject has purposes, goals, beliefs, desires, motivations, values, emotions or intentions that genuinely 'make a difference'—*qua* intrinsically meaningful states (i.e., in virtue of their meaning or content)—to how the agent behaves.[8]

This is reflected in the phenomenology of agency. A fundamental aspect of feeling that our actions are 'up to us' is the perception that we are acting based on what is meaningful to us: our motives, concerns, or values. We seem to act *for reasons*, which we consider to be both motivating and intelligible. By contrast, the jerking movement you make when a doctor taps your knee feels like it was not 'up to you', precisely because no belief, desire, or intention seemed to play any part in causing it.

This connection between meaningfulness and action is not only reflected in our lived experience of agency; it also underpins the normative structure of the concept. For an action to be genuinely attributable to an agent—such that they can be praised, blamed, or otherwise held accountable for it—the action must be the kind of thing that can be evaluated in terms of success conditions, rational standards, or moral norms. Only when an action is truly performed *for* reasons—when it is the product of an agent's motivations and subjective perspective—can it (non-metaphorically) be described as something that was successful or unsuccessful, right or wrong, justified or unjustified, rational or irrational. And this kind of evaluability only intuitively makes sense in the context of semantic causation, that is, only when an agent's behaviour *can* be shaped by what is intrinsically meaningful to them. As Susan Wolf (1990) notes, "only an agent who has a will—that is, who has desires, goals or purposes, and the ability to control her behaviour in accordance with them—can be responsible for anything at all" (p.7).

For Froese and Taguchi (2019), any naturalistic account of biological agency must therefore provide a credible scientific understanding of semantic causation, which they describe as "a theory of the living that makes room for meaning and intentional action to make a difference on their own terms in the natural world" (p.4). Only then could one justifiably describe an organism, or any system, as being "capable of adaptive, goal-directed behavior" (Budaev et al., 2019, p.14), as having "the capacity to act for a goal or

[8] In the literature, this is often referred to as mental causation. I prefer the term 'semantic causation' because it foregrounds the more basic notions of content and meaning, rather than leaning on the more metaphysically loaded notions of intentionality, mind and consciousness. In doing so, the term allows for a broader conception of meaningful causation which is not restricted, by definition, to more complex, representational or conscious agents, but may also be applicable to simpler systems too.

purposes guided by norms" (Fulda, 2016, p.ii; see also Walsh, 2015), as able to "act on their own behalf" (Jaeger, 2024, p.169, see also Kauffman, 2000; Moreno & Mossio, 2015, ch.4), or as having "the capacity to act intentionally" and "for reasons" (Schlosser, 2019)—all of which are standard ways of defining agency in the literature.

The problem, however, is that the notion of semantic causation has typically been treated with a great deal of naturalistic suspicion. As Froese and Taguchi (2019) also note, "the perennial philosophical problem of how to make room for subjective meaning in a physical world, first articulated in its modern form by Descartes, remains fundamentally unsolved" (p.1). Broadly, the Cartesian worry here is that the world as it is given to us by natural science seems to be what became known during the Enlightenment as a 'clockwork universe': a world of purely physical cause-and-effect, where everything that happens *just happens* as the consequence of physical interaction or law—with no room for anything like meaning or purpose within that causal schema (Riskin, 2016). This is sometimes known as the 'challenge from eliminative materialism' (Churchland, 1981).

In developing his influential Causal Exclusion Argument, Jaegwon Kim (1998, 2005) has perhaps done more than anyone to advance this challenge for semantic causation (and, by extension, agency). In that argument, Kim's central claim is that if every physical event already has a complete and sufficient physical cause—as the clockwork universe suggests—then there is simply no causal work left for the intrinsic meaning of subjective or mental states, such as beliefs and desires, to do. For these states to genuinely 'make a difference' to behaviour, their causal efficacy must ultimately be derived from their physical substrates—neuronal activity, for example. The alternative—i.e., "for meaning and intentional action to [also] make a difference on their own terms" (Froese & Taguchi, 2019, p.4)—would create an apparently problematic form of 'causal overdetermination', wherein a single physical effect would have both a fully sufficient physical cause and an additional, irreducible semantic cause (Kim, 2005). On this view, semantic (or mental) content becomes ultimately *epiphenomenal*—explanatorily redundant artefacts of neural activity that are insufficient to ground the claim that agents genuinely act in virtue of their reasons or intentions.

This sort of skepticism toward semantic causation has been claimed to find empirical support within contemporary behavioural and neuroscientific research. Robert Sapolsky (2023), for example, surveys a wealth of evidence demonstrating how genetic and environmental factors can help shape who we become and how we behave, as a means of arguing that human behaviour is effectively the product of automated biochemical and neurophysiological processes. Similarly, Daniel Wegner (2002) extensively examines the

phenomenon of confabulation in humans, concluding that we are often not even aware of our own motivations, let alone acting on the basis of their intrinsic content. Perhaps most prominently, Benjamin Libet and colleagues (1982) present experimental findings that appear to show neurophysiological markers of a subject's decision that reliably precede the subject's own conscious awareness of the decision, which many have interpreted as indicating that conscious decision-making is not, in fact, causally efficacious after all. Collectively, these lines of empirical research have been widely interpreted as corroborating Kim's philosophical challenge to semantic causation.

If this is right, then "the upshot is epiphenomenalism; [an] agent's intentions and other mental states seem causally inert; they have no genuine manifestation" (List, 2019, p.46). The semantic content of our mental states—the meaning and purpose we experience in acting—would be, at best, *effects*, not causes. They would be the exhaust fumes given off by the neurophysiological processes in our brains and bodies that *actually* do the causal work in producing our actions. The result would be that semantic causation *is* therefore fundamentally incompatible with the scientific worldview. In turn, this would mean that appeals to agency—with its metaphysical commitment to genuine semantic causation—would be little more than appeals to some sort of supernatural or mysterious forces, an *entelechy* or *elan vital*.

In this thesis, I will therefore follow Froese and Taguchi (2019) in assuming that one of the key challenges for a naturalistic theory of agency is to overcome Kim's argument against semantic causation by developing an empirically grounded "account for how meaning as such could make a difference for an agent's behaviour" (p.1).

Indeed, it is worth pausing here to note that, of the challenges facing naturalistic accounts of agency, this one has probably garnered the most attention among existing theories. The theory of Situated Darwinism, for example, developed by philosopher Denis Walsh (2015) as a challenge to the mainstream Modern Synthesis approach within evolutionary theory, advances what is arguably the most influential contemporary account of biological agency. In it, Walsh seeks to naturalise the purposive dimension of agency by appeal to a certain kind of explanatory strategy (cf. Fulda, 2017). According to this view, if the behaviour of a macroscopic entity exercising agent-level causal control can be accurately described by citing a particular goal, purpose or reason to which the behaviour conduces (i.e., if the behaviour can be modelled teleologically) then that just *is* sufficient evidence to describe the behaviour as an exercise of the organism's agency. For example, an immune cell following and engulfing a bacterium counts—on this view—as an exercise of the immune

cell's agency because we can explain this behaviour by citing a goal to which it reliably conduces (namely, to catch the bacterium).

In **Chapter 6**, I argue extensively that Situated Darwinism seems to sidestep, rather than address, the challenge posed by eliminative materialism for the semantic causation aspect of agency. Rather than requiring agential behaviours to follow from a particular kind of internal state (goals, purposes, beliefs), this approach to naturalising agency simply allows the organism's 'goal' to be in the eye of the observer, so to speak. No account is given for "how meaning as such could make a difference for an agent's behaviour" (Froese & Taguchi, 2019, p.1). Instead, the organism is, to some extent, black-boxed—and, hence, one might worry that this account of agency actually just explains *away* semantic causation rather than providing a robust naturalised understanding of it.

Another extant attempt to naturalise agency comes from a compelling theoretical framework for understanding and explaining biological organisation and autonomy (Moreno & Mossio, 2015; see also Barandiaran et al., 2009). This theory of autonomy advances a very similar account of agency to that of Situated Darwinism, but with one important addition. It takes seriously the problem of semantic causation, by attempting to ground the system's motivational state (its goals, purposes, or reasons) in its intrinsic dynamics.

The theory itself is richly developed and highly technical, making a full exposition here beyond the scope of this introduction. However, the general idea is that living systems are observably distinguishable from non-living systems in virtue of a particular kind of organisational architecture. This organisational architecture is defined by the observation that all of the elements in the living system reciprocally depend on and produce one another; they must all perform their function in order to continuously generate the whole system of which each element is a part. The important point then is that, collectively, these dynamics are both necessary and sufficient to create the conditions for the system's own continued existence. That is, living systems instantiate the Kantian ideal in which "the parts exist for and by means of the whole, and the whole exists for and by means of the parts" (Kauffman, 2013, p.609). The authors of the theory claim that this means the system is rightfully thought of as 'self-determining': the collective activity of the whole, integrated system is what establishes and determines the system's own continued persistence, by continuously generating and maintaining the conditions necessary for its own ability to resist entropy and continue persisting. In other words, it realises a sort of organisational *closure* (Moreno & Mossio, 2015; Jaeger, 2024).

The philosophical consequence of this, for my purposes, is that there appears to be an intrinsic purposiveness to the organism's activity. Everything it does (and everything its parts do) is oriented toward creating the conditions for its own ongoing survival. The system can therefore be seen as exerting effects, holistically, *with the purpose of* survival.

There is thus an inherent normativity to this. Certain effects (hence, certain behaviours) will be 'better' than others with respect to the goal of self-persistence. The system can then learn to adapt its behaviours to improve the chances of bringing about these effects, thereby establishing some sort of tangible link between the motivational state of the system (its intrinsic purposiveness) and its behaviour. As some of the theory's authors explain:

> "The theory of autonomy offers, therefore, a perspective from which agency can be understood as behavior performed for a reason, directed towards an intrinsic goal, which is the continued existence of the system's self-determining organization" (Virenque & Mossio, 2023, p.2)

This approach therefore offers one of the most promising naturalistic foundations for understanding semantic causation and purposive agency. It provides an empirically grounded account of how causal power could inhere at the level of the whole system's organisation (macroscopic causal control) and involves an intrinsic motivational state that appears to 'make a difference' to the dynamical behaviour of the system (semantic causation).

However, it is not clear whether this account, on its own, has the resources to naturalistically account for the sort of semantic causation at issue in the more complex forms of biological agency I am interested in in this thesis.

There are two main reasons for this. First, as Alex Djedovic (2020) has argued, by tethering agency entirely to a goal of persistence, this approach appears to "foreclose[…] on the possibility of agents having goals that normatively require actions that are neutral for persistence or deleterious to it" (p.74). Not only is this incongruent with our phenomenology of agency—in that, it rarely feels like we are acting explicitly 'in order to survive'—it also appears insufficient to ground the claim that agents act on the basis of the sorts of reasons and motives for which we typically hold people responsible. We do not generally praise or blame people for succeeding or failing to meet their goal of continued existence; we judge them for acting on the basis of much 'higher level' goals and reasons, such as keeping a promise to a friend or supporting a political candidate's economic stance. It also seems to be incompatible with the idea that agents make genuine choices

about how to act (as we will discuss in **Section 1.2.3.**). Organisms presumably have no control over their 'goal to persist', this is simply a statistical tendency that is conferred on them by natural selection. Thus, if this 'goal' is what informs and drives their behaviour, then it does not appear as though the organism has real 'say' in the matter. Its behaviours are instead foisted upon it by the demand to ensure ongoing persistence, a demand that it did not itself issue.

The second reason to think this account of agency may not be sufficiently robust to meet the conceptual demands of a natural theory of biological agency is that it is not clear that the conditions for semantic causation are actually being met here. Specifically, it is not clear that the agent's intrinsic goal or purpose is *really* 'making a difference' to what the system is doing on this account. As Froese and Taguchi (2019) explain:

> "while at first sight the notion of "behavior according to an intrinsic norm [or goal]" suggests that an agent behaves the way it does because that is what is good for it, it is actually more correct to say that it behaves that way simply due to certain dynamical constraints on its internal and interactional dynamics. Whether we label a region of this dynamical state space as being in line (or in tension) with its norms [or goals] simply makes no difference for the agent, nor to the unfolding of its trajectories. Once the agent has started on a particular trajectory, it can never leave that trajectory—even if it is trajectory [sic] that is supposedly intrinsically bad for it. To repeat, in this framework the norm as such makes absolutely no difference to activity: the behavior is something that is just undergone by the system, rather than actively chosen to be in accordance with an intrinsic norm [or goal]." (pp.3-4)

Consequently, in this thesis, I pursue a more robust account of semantic causation. In particular, I argue for a form of informational causation (in neural systems, at least) wherein patterns of neural/physical activity have causal influence within the system *in virtue of what they mean for the system* (**Chapter 3**). That is, where it is primarily the *meaning* of the patterns that is causally efficacious in bringing about a behaviour, rather than the physical activity *per se*. This allows us to model semantic content and subjectivity not merely as epiphenomena, but as genuinely causal aspects of biological agents.

1.2.3. Genuine Choice and 'the ability to do otherwise'

The third feature of our intuitive concept of agency that I will focus on in this thesis is the notion of genuine choice—the idea that it is open to an agent to choose among alternative courses of action. On this view, agency is not *only* about purposive, agent-level control.

There also needs to be some sort of freedom or real-time discretion involved in *how* one exercises that control. In everyday life, for example, the sense is often that our actions are 'up to us' precisely *because* we could have done something else—we acted in a way that we "need not have done" (Steward, 2012, p.164).

As Peter van Inwagen (1983) notes, this idea of having options to choose between is already implicitly presupposed every time we deliberate about something. He writes: "it seems to be a feature of our concept of deliberation that we can deliberate about which of various mutually exclusive courses of action to pursue only if we believe that each of these courses of action is open to us" (p.30). It also seems to be built into the functional role that the concept of agency plays for us in our practices of moral and legal responsibility. In most cultures, we blame a wrongdoer because we believe he *should* have chosen a better path—a judgement that would seem to lose its meaning if in fact he *could not* have chosen differently. (Indeed, this is the intuition behind the Kantian principle that "ought" implies "can." (Kant, 1781/1929)). Conversely, we also tend to excuse or mitigate responsibility in cases where a person appears to have *lacked* genuine choice, as illustrated by certain insanity or duress defences in criminal law. Thus, part of what we take it to mean for an outcome to be 'up to us' appears to be "inextricably bound up with the idea that we are capable of making real choices, at least in principle" (List, 2019, p.2)

Notably, this idea—of having alternative possibilities and making genuine choices between them—has been at the heart of philosophical discussions of free will for millennia. Since at least the time of Epicurus and Aristotle, theorists have argued that the sort of agency characteristic of free will demands what is known in this literature as a sort of 'leeway freedom' or the 'ability to do otherwise' (see Timpe, 2016; O'Connor & Franklin, 2022). Prominent free will philosopher Robert Kane (1996; 2007; 2019; 2024) has also referred to it as an agent having 'plural voluntary control' over what they do: "they have the power to act voluntarily, intentionally, and rationally in more than one way, rather than in only one way" (2024, p.19). Libertarians about free will, like Kane, hold to the view that such genuine choice (or plural voluntary control) is both required for an agent to have free will *and* that we, as human agents, possess this ability.

To be sure, the conceptual necessity of alternative possibilities and genuine choice is certainly not universally accepted among free will theorists: compatibilist philosophers, for example, argue that an agent could still act freely and/or be held responsible even if it was never physically possible for them to *actually* 'do otherwise' (Frankfurt, 1969; Fischer & Ravizza, 1998; Dennett, 2015). In this thesis, my interest is more in what kinds of freedom or agency we might plausibly *have*, rather than the question of which kinds of

freedom are conceptually necessary—or "worth wanting" (Dennett, 2015)—to justify moral responsibility practices. Thus, I follow (without much argument) Steward (2012) and colleagues who contend that, not only is genuine choice required for (libertarian) free will, it is central to our notion of *agency* itself. For the purposes of this thesis, I will therefore treat genuine choice as one of the core aspects of our intuitive concept of agency, one that any naturalistic account will need to accommodate.

This is easier said than done, however. The notion of genuine choice carries with it several metaphysical commitments that make it particularly challenging to reconcile with the natural sciences. As Robert Kane (2024) explains, it seems to imply both "that (i) the future is open, with multiple [physically] possible paths into the future, and (ii) it is sometimes "up to us," and no one and nothing else, which of these possible paths will be taken" (p.2). Each of these conditions, however, have been the subject of their own long-standing naturalistic critique.

First, for the future to be genuinely open in the manner required for genuine choice, it is natural to think there must exist multiple "real possibilities" in our universe, where "what *is really possible* in a given situation is what can temporally evolve from that situation against the background of what the world is like" (Müller et al., 2019, p.3). I cannot be genuinely choosing between doing A or B if, in reality, only A is physically (i.e., *really*) possible.

Yet for much of the modern scientific era, the prevailing view of nature has been one of determinism, where "there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.3). The idea, in effect, is that the "laws of physics determine the future" (Greene 2013, as quoted in Mitchell, 2023a, p.11). On this view, atoms are simply bumping into each other, according to deterministic physical laws, generating a relation of logical necessity between the complete state of the universe at time *t* and its subsequent state(s) at *t+n*.

From the early Greek atomists, to Newton's discovery of his laws of motion, to Laplace's evocative image of a demon who could predict everything that will ever happen, and even to more contemporary stimulus-response models of behaviour, scientists have often portrayed our universe in this way: as every event (including human decisions) following inevitably from the events that precede it. On this view, the future is fixed in principle by the past and the laws of nature.

This poses an obvious challenge for any account of agency that treats genuine choice as essential (e.g., libertarian free will): if only one outcome is ever physically possible, then

any talk of "options" or "choosing otherwise" becomes merely an illusion. As van Inwagen (1983) explains in his highly influential Consequence Argument:

> "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us." (p.56)

A naturalistic account of biological agency must therefore grapple with the problem of how alternative possibilities *can* exist both in the physical world generally, and in biological decision-making and action selection processes in particular. This is the classic 'challenge from determinism'.

In this thesis, I address this challenge by arguing that the empirical evidence from physics does *not*, in fact, support the view that determinism holds true in our universe, at any level of description. On the contrary, I argue that the evidence favours a fundamentally *in*deterministic worldview instead—wherein the inherent indeterminacy is of a sort that is especially salient for complex, chaotic systems like us. The upshot is that it very much *is* the case, for reasons independent of free will itself, and for biological systems especially, that "the future is open, with multiple [physically] possible paths into the future" (Kane, 2024, p.2)—as is required by the notion of genuine choice.

Embracing indeterminism in this way runs us straight into a second naturalistic concern, however. That is because, even if we grant that the future might be open in the requisite way, genuine choice demands more than just having multiple physically possible paths available to us—it requires that "it is sometimes "up to us," and no one and nothing else, which of these possible paths will be taken" (Kane, 2024, p.2). The existence of alternative possibilities is therefore necessary but not sufficient for the sort of agency or freedom at issue here; the agent must also somehow exercise control over *which* possibility gets realised. This is where the challenge from indeterminism emerges, often known in the philosophical literature as the "Luck Objection" (see Mele, 2006).

Proponents of the Luck Objection (e.g., Haji, 2004; Levy, 2011; Pereboom, 2014) argue that embracing a metaphysical picture of indeterminism actually serves to undermine (rather than secure) our capacity for genuine choice because all it does is introduce an element of randomness into the decision-making process, which—it is argued—renders it a mere matter of chance or luck, for the agent herself, how this process will play out. The worry is that:

> "If the occurrence of a choice depends on the occurrence of some undetermined or chance events (e.g., quantum events) in the brain over which the agent lacks control, then whether or not the choice occurs would appear to be "just a matter of luck" rather than something the agent had control over and was responsible for." (Kane, 2019, p.154)

Hence, the notion of genuine choice (or the ability to 'do otherwise') seems to come with a built-in dilemma. On one horn, if our actions are determined, then it is (supposedly) possible for us to have control over what we do but there will be no genuine alternatives for us to choose between. On the other horn, if our actions are not determined, then we gain the alternatives required for genuine choice but we risk losing control (since an undetermined choice might just be a random event). This dilemma forms a formidable challenge for any sort of scientific understanding of genuine choice: how can we have real options available to us that are neither precluded by physical law nor the product of mere randomness?

Reconciling genuine choice with scientific naturalism requires navigating between these two threats. It demands a picture of the universe in which (i) the future is genuinely open in at least some situations, and (ii) an agent can reliably settle which of the possible futures comes about, in a way that is not just a matter of luck or chance. In this thesis, I argue that such a picture is not only possible but is actually supported by contemporary science. As mentioned, I will argue that current physics gives us reason to believe that strict determinism is false—the world is not a pre-written script. In doing so, however, I will also argue for an alternative *conception* of indeterminism to the one that is traditionally deployed in these debates. This alternative conception of indeterminism, I will suggest, supports a model of decision making in biology that is both indeterministic *and* under the agent's control. The argument, therefore, is that agents *can* have real alternative possibilities *and* be the genuine source of which possibility becomes reality. In this sense, then, I hope to develop a naturalised account of the idea of genuine choice that sits comfortably within a scientifically grounded understanding of human agency.

Moreover, in doing so, I also aim to offer a defence of libertarian free will. As already mentioned, the libertarian view of free will is effectively defined by its commitment to the notion of genuine choice (or 'the ability to do otherwise'). Consequently, it is often regarded as scientifically implausible and even somewhat naïve. By developing a novel conception of indeterminism and defending the notion of genuine choice on that metaphysical basis, however, this thesis aims not only to overcome these challenges for

libertarianism, but also to offer an alternative *account* of libertarian free will to the ones that these challenges typically target.

1.3. Summary

In this introduction, I hope to have demonstrated the significance that a more naturalistic understanding of agency could potentially have for how we perceive ourselves, how we structure our societies and our interpersonal dynamics, and how we conduct research in the biological sciences and the field of AI.

I then outlined three specific challenges for a naturalistic account of agency. These challenges will be the primary focus of this thesis. To summarise, they are:

(i) **The challenge from reductionism for causal sourcehood:** A challenge to the idea that agents are the causal source of their actions, based on an argument that natural science implies a causally reductive worldview that, in one way or another, precludes biological systems from exhibiting the sort of emergent, macroscopic, agent-level causation that many (including myself) consider to be essential for genuine sourcehood, and thus agency.

(ii) **The challenge from eliminative materialism for semantic causation**: A challenge to the idea that an agent's subjective meaning and intentional states have causal efficacy, *in virtue of their meaning or content*, in bringing about action. This challenge is based on the argument that, under a naturalistic worldview, such mental or semantic states are, in fact, mere epiphenomena—exhaust fumes given off by the biochemical processes that *actually* cause the behaviour.

(iii) **The challenge from determinism *and* indeterminism for genuine choice:** A challenge to the idea that agents genuinely have the power to choose how to act from among multiple different options. This challenge is based on a two-pronged argument which states that either (a) natural science supports the metaphysical thesis of determinism and, hence, there are never any *real possibilities* available to the agent to choose between, or (b) the sort of indeterminism that physics suggests *may* exist in our universe provides agents with the *real possibilities* required for genuine choice, but only by introducing randomness and chance into the equation which strips agents of the control required to *choose* which of the real possibilities becomes actual.

Collectively, these three challenges form a major part of the reason why—despite there being strong phenomenological, academic, and socio-cultural reasons to *expect* that agency is a fundamental feature of living systems—the concept continues to lack a firm scientific grounding.

The aim of this thesis is to contribute to such a grounding. To do this, I approach each challenge through the lens of a series of existing debates in the philosophy of action (**Chapter 2**), behavioural neuroscience (**Chapter 3**), the philosophy of free will (**Chapters 4** and **5**), and evolutionary developmental biology (**Chapter 6**)—each of which centres around one or more of the relevant features of agency mentioned above.

An important point to note, however, is that these features—and their respective challenges—are typically not treated independently from one another, in the way I have been presenting them here. In reality, these lines of naturalistic critique are often tightly interconnected: skeptics almost always combine, mix (and sometimes even conflate) them in making their arguments. Consequently, debates about agency in the literature rarely cover just one of these challenges or features of agency in neat isolation. To give an example, the concept of agent causation—as it is discussed in the literature—is often not taken to be concerned *solely* with the existence of emergent, macroscopic causation (as I have been suggesting for the purposes of exposition). Part of the discussion also tends to centre around agents being the *initiator* or *originator* of their actions (Kane, 1996; O'Connor, 2000; Clarke, 2003). This therefore brings in issues related to the 'challenge from determinism': how can an agent—even one exhibiting agent-level causation—be the true *originating* source of an action, if they were physically compelled to perform that action by the laws of nature, the state of the environment, and/or their genetic inheritance? As McKenna (2003) notes, the 'challenge from determinism' is therefore *also* relevant to the problem of naturalising causal sourcehood (in addition to the 'challenge from reductionism') because the truth of determinism would seem to entail that "the sources of an agent's actions do not originate in the agent but are traceable to factors outside her" (p.201).

Similarly, as implied by the discussion in **Section 1.2.3.**, to successfully address the Luck Objection to genuine choice, one ultimately needs to show how agents in an indeterministic universe can still be 'in control of' their decisions and actions, despite the presence of indeterminacy and chance in their decision-making processes. To demonstrate such control, however, requires giving an account of how the conditions for *causal sourcehood* and *semantic causation* could plausibly be met within an indeterministic universe (the topic of **Chapter 5**).

The point then is that, in order to properly engage with these challenges and their respective literatures, it is not feasible for me to address each line of attack in isolation. To do so would be to over-simplify the problem and, hence, to not engage productively with the relevant literature. Consequently, readers should be aware that this thesis is *not* intended to be structured in such a way that each substantive chapter of the thesis is strictly dedicated to addressing a *single* challenge from naturalism or to naturalising a *single* feature of agency. Broadly, **Chapters 2** and **3** are intended to *collectively* address the threats to agent-level causation and semantic causation (with **Chapter 2** primarily focusing on agent-level causation and **Chapter 3** primarily focusing on semantic causation). Yet both chapters include extensive discussion of determinism and indeterminism. Likewise, **Chapters 4** and **5** are explicitly aimed at addressing the two threats to genuine choice—first from determinism (**Chapter 4**) and then from indeterminism (**Chapter 5**). But **Chapter 5** in particular discusses both macroscopic control and semantic causation at some length. This thesis is therefore structured such that the argumentative load for addressing the three naturalistic challenges presented above is distributed across the whole thesis, rather than on a strict chapter-by-chapter basis. As such, readers should be aware that there is some degree of crosstalk and conceptual overlap between the chapters.

Some further caveats are also in order. First, I want to note that, in defending the naturalistic credentials of these three aspects of agency, I do not consider myself to have attempted a wholesale defence of the intuitive concept of agency from the threats of scientific naturalism. There are several other features of agency that I consider to be a part of our intuitive understanding of the phenomenon that I will not address in this thesis. For instance, I will not address the notion of ultimate responsibility in which, for our choices and actions to be 'up to us' in such a way that we can be justly held morally and legally accountable for them, we must somehow be responsible for becoming the *type of person* that we are, in that moment (Kane, 1996, 2024; Levy, 2011; Harris, 2012; Sapolsky, 2023). Similarly, I will not address the problem of consciousness and the idea that agential choices and actions must be *consciously* caused in some way (Libet et al., 1983; Wegner, 2002; Harris, 2012; Levy, 2014; Shepherd, 2015). While I take both of these to be interesting and potentially important forms of agency, they are beyond the scope of this thesis.

The second caveat I should give is that, in engaging with this project, I am not intending to imply (or to defend) the view that *all* types and conceptions of agency are ultimately going to be naturalisable. In other words, in offering an empirical defence of causal sourcehood,

semantic causation, and genuine choice, I am not also committing to the view that alternative conceptions of agency—such as Nietzsche's notion of a "causa sui" (1886, Section 21, see also Strawson, 1994) or Chisholm's "prime mover unmoved" (1964, p.9)—might also prove to be consistent with modern science. Instead, this thesis should be seen as a targeted contribution within a much broader research programme. My aim is only to address the three specific aspects of agency introduced in this section, largely leaving open the question of whether other conceptions may also be reconcilable with a naturalistic metaphysics.

Moreover, it should be noted that this thesis has only a very small amount to say about the specifically *conceptual* questions surrounding agency—that is, questions about what should and should not be included as a feature of one's *concept* of agency (or what should and should not be considered a *type* of agency). My interests here revolve more around which of the *potential* features/types of agency represent naturalistically plausible properties of biological systems like us, and which do not. I am less interested in, as Daniel Dennett (2015) put it, which sorts of freedoms or types of agency are "worth wanting".

Finally, I would also like to acknowledge that each chapter of this thesis was written as a standalone article that could be published in a philosophical or scientific journal. As such, the chapters can be read independently and in any order (although I suggest reading them in the order in which they appear since, to my mind, this provides the most logical progression of the ideas). It also means, however, that there is some repetition from chapter to chapter. For example, the empirical argument for *in*determinism is presented multiple times and the different types of causal reductionism are (re-)introduced in several places. Furthermore, because each article was written to be situated within specific (and somewhat diverse) literatures, the terminology and explicit focus of the chapters varies slightly from the overarching themes set out in this introduction. To remedy this, in the next section I outline each chapter's role in the thesis's overall argument, and in the Discussion (**Chapter 7**) I explicitly synthesise and summarise the contributions of each chapter to the overall themes of the thesis (i.e., causal sourcehood, semantic causation, and genuine choice).

In the next section, I give a more detailed overview of each chapter.

1.4. Chapter Outline

All of the chapters in this thesis have either been previously published or are currently under review. They are all based on co-authored work, for which I was the lead author.

In this section, I provide an overview of each chapter, including the specific research question that the chapter aims to address and a summary of the arguments therein.

Chapter 2 | Naturalising Agent Causation

Primary Research Question: Can the notion of macroscopic, agent-level causation be reconciled with modern empirical science, at least in principle?

Summary

The notion of agent causation—i.e., that an entity, such as a living organism, can be a cause of things in the world *in its own right*, rather than merely as a consequence of events and processes going on *within* it—is generally considered to be mysterious and fundamentally incompatible with our best scientific theories. Here, I argue against this view. I present eight empirical criteria which, I contend, collectively describe a naturalistic theoretical system that exhibits an entirely non-mysterious form of agent causation. The criteria are: (1) thermodynamic autonomy, (2) persistence, (3) endogenous activity, (4) holistic integration, (5) low-level indeterminacy, (6) multiple realisability, (7) historicity, (8) agent-level normativity. I show how any system or entity that meets these criteria can overcome the apparent naturalistic challenges facing agent causality, at least in principle. I also suggest—but do not definitively defend—the view that, in practice, many living systems *do* meet these criteria. This chapter therefore lays the theoretical groundwork for later chapters in the thesis to demonstrate that biological systems really do satisfy these criteria in practice and, thus, exhibit agent-level causal power.

Chapter 3 | Beyond Mechanism: Extending Our Concepts of Causation in Neuroscience

Primary Research Question: Is the activity of localised, synchronic neural mechanisms sufficient to provide a complete causal explanation of why a particular behaviour occurred?

Summary

Neuroscientists often search for the neural mechanisms of behaviour: individual neurons, neural circuits, or neural populations whose activity seems to causally "drive" a behaviour of interest into effect. With optogenetic technologies, researchers can now effectively exert exogenous control over many animal behaviours, selectively initiating (or inhibiting) them simply by (in)activating isolated sets of neurons within the animal's brain. This creates the

strong impression that, even under more naturalistic conditions, it is ultimately the synchronic activity of these localised neural mechanisms that is *causing* the particular behaviour to occur, such that one could provide a sufficient causal explanation of the behaviour simply by citing those synchronic neural mechanisms. The apparent upshot for the philosophy of agency is that these systems are being 'pushed around by their parts' in a way that leaves no room for agent-level causation or semantic causation. Here, I argue that this is not the right way to think about the neural causes of behaviour. I develop and defend a more expansive view of causation within the brain, by integrating recent neuroscientific evidence with a range of established causal concepts from across the philosophy of causation. These include concepts of criterial causation, constraints, triggering and structuring causes, macroscopic, informational causation, and historicity. Under this causal schema, synchronic neural mechanisms provide only a partial answer to the question of "what caused this behaviour to occur?". In addition, I argue, one must also take into account non-reductive, temporally extended, and meaning-laden forms of causation if one is to provide a complete causal explanation of an animal's behaviour. This chapter therefore offers a naturalistic account of how both agent-level causation and semantic causation are realised, in practice, within neural systems.

Chapter 4 | Reframing the Free Will Debate: The Universe Is Not Deterministic

Primary Research Question: Is the idea that it is physically open to us to take more than one possible course of action compatible with contemporary physics?

Summary

In this chapter, I turn to the notion of genuine choice. In the philosophy of free will, a common argument against the existence of genuine choice (or 'the ability to do otherwise', as it is often referred to in this literature) centres around the metaphysical thesis of determinism. Many contend that, if determinism is true, then agents are not *really* free to choose how they behave because such choices require a physically open future and determinism entails a physically closed future. Here, I survey a range of empirical and conceptual arguments from across modern physics to argue that our best scientific theories strongly support the view that determinism does *not* hold at any level of description in our universe, and especially not for non-linear, macroscopic systems such as ourselves. I then explore the implications of this for the current theoretical landscape of the free will debate. Most notably, I show that these arguments from physics apply equally

to the sorts of *in*deterministic worldviews typically discussed by free will philosophers (a model of indeterminism I call 'determinism-plus-randomness'). I develop a novel and empirically grounded, *alternative* concept of indeterminism—'pervasive indefiniteness'—which I propose as the primary metaphysical starting point for discussions about the existence of free will. This chapter therefore defends the notion of genuine choice from the 'challenge from determinism' by arguing that, from the perspective of scientific naturalism, it is highly unlikely that there *is* such a threat. It then lays the groundwork for a novel defence of genuine choice from the 'challenge from luck' by motivating and developing an alternative *conception* of indeterminism to the one that standardly underlies this challenge.

Chapter 5 | Chance, Choice, and Control: Free Will in an Indeterministic Universe

Primary Research Question: Can the notion of genuine choice be reconciled with a modern scientific understanding of deliberative decision-making?

Summary

The second horn of the standard argument against genuine choice is the 'challenge from luck'. On this view, agents in an indeterministic universe do not have the freedom to choose how they will behave, not because there are no physically possible alternative futures available to them, but because they lack the control required to select which of these possible futures becomes actual. It is argued that the sort of indeterminism that might exist in our universe (quantum indeterminacy) only serves to introduce an element of randomness into our decision-making process, over which we have no control, thereby making our 'decisions' and actions mere matters of chance or luck, and not something we have the power to actively choose. Here, I integrate the conclusions from previous chapters to defend the notion of genuine choice (and libertarian free will more generally) against these so-called Luck Objections. I do so on two fronts. First, I consider and critique five prominent versions of the objection, showing that the majority of these arguments rely on assumptions and claims about indeterministic decision-making processes that we have good empirical reason to reject. Second, I use these negative arguments to develop a positive, alternative, and more empirically plausible model of deliberative decision-making in humans. Under this model, I argue, our decisions are neither pre-determined nor random; they are the outcome of an under-determined deliberative process over which agents exercise macroscopic, reasons-guided, constraint-based control. This chapter

therefore completes the naturalistic defence of genuine choice started in **Chapter 4**, by applying the accounts of agent-level causation and semantic causation developed in **Chapter 2** and **3** to the context of indeterministic deliberative decision-making. The result, I suggest, is an entirely naturalistic account of human decision making in which it is both (i) open to us, prior to deliberation, to take more than one action, and (ii) 'up to us', during deliberation, which of those possible actions we take.

Chapter 6 | A Critique of the Agential Stance in Development and Evolution

Primary Research Question: Is the claim that 'developing organisms exercise a form of agency over their own development which cannot be explained by contemporary evolutionary theory' really consistent with the empirical evidence?

Summary

In this final chapter, I turn to some recent attempts to (i) naturalise agency within the context of developmental biology and (ii) use this as motivation for a radical re-formulation of evolutionary theory. These accounts claim that empirical research in developmental biology supports a view in which developing organisms actively and purposively control their own development toward adaptive outcomes. The apparent upshot, it is claimed, is that we ought to re-conceptualise biological evolution primarily "as the consequence of organisms' pursuit of their goals" (Walsh & Rupik, 2023, p.10). Here, I offer a critique of this agential perspective on development and evolution. I show that, in order to ground the radical implications for evolutionary theory proposed in (ii), the claims about purposive developmental agency in (i) need to be interpreted literally. However, I argue that a literal interpretation is not well supported by the wider empirical evidence base within developmental biology, despite claims to the contrary. Hence, I conclude that there is insufficient evidence for a radical re-formulation of evolutionary theory (at least on the basis of the sort of naturalistic developmental agency at play in these accounts). This chapter therefore takes a critical lens to the wider project of naturalising agency of which this thesis is a part. It once again asks: are claims about biological agency compatible with scientific naturalism? But, this time, I come down on the side of skepticism.

Chapter 2

Naturalising Agent Causation

Status

This chapter has previously been published as:

> Potter, H. D., & Mitchell, K. J. (2022). Naturalising Agent Causation. *Entropy, 24*(4), 472. https://doi.org/10.3390/e24040472

Author contributions

Equal contribution—both authors conceived, wrote, and edited the manuscript together.

2.1. Abstract

The idea of agent causation—that a system such as a living organism can be *a cause of* things in the world—is often seen as mysterious and deemed to be at odds with the physicalist thesis that is now commonly embraced in science and philosophy. Instead, the causal power of organisms is attributed to mechanistic components *within* the system or derived from the causal activity at the lowest level of physical description. In either case, the 'agent' itself (i.e., the system as a whole) is left out of the picture entirely, and agent causation is explained away. We argue that this is not the right way to think about causation in biology or in systems more generally. We present a framework of eight criteria that we argue, collectively, describe a system that overcomes the challenges concerning agent causality in an entirely naturalistic and non-mysterious way. They are: (1) thermodynamic autonomy, (2) persistence, (3) endogenous activity, (4) holistic integration, (5) low-level indeterminacy, (6) multiple realisability, (7) historicity, (8) agent-level normativity. Each criterion is taken to be dimensional rather than categorical, and thus we conclude with a short discussion on how researchers working on quantifying agency may use this multidimensional framework to situate and guide their research.

2.2. Introduction

When an organism acts in the world, is it right to say that *the organism* caused the effect? Or is that simply a useful metaphor or a convenient level of description? Perhaps the more accurate statement is that some biochemical pathway or neural activity *within* the organism caused the effect. In other words, is it right to think of organisms as agents capable of action? Or are they simply loci of complicated happenings that give the illusion of concerted, autonomous agency? Are there things that happen in the world that are rightly said to be "up to" organisms, things that *they do* as agents, or are such happenings, in fact, reducible to the physical evolution of the components that constitute the organism? These are the questions posed by the concept of agent causality.

Introduced by Thomas Reid in 1863, "agent causation" is the claim that agents themselves can be the cause of events (Reid, 1983). That is, organisms or systems *as a whole* can have causal power that is not entirely reducible to the causal power of their component parts or determined by the states of the environment (Markosian, 1999; Pereboom, 2004; O'Connor, 2009; Steward, 2012). Therefore, it mirrors a much older concept, framed by Aristotle and Epicurus in the context of human beings, of some things being "up to us".

To justify agent causation, then, systems need to exhibit a causal power that *irreducibly* inheres at the level of the whole system (i.e., it is nonreductive) while still maintaining that the causal power is instantiated in, or realised by, the system's physical constituents (i.e., it obeys physicalism). Timothy O'Connor (2009) describes this as an "ontologically primitive causal power" that is "at once causally dependent on microphysically-based structural states and yet ontologically primitive" (p.195)—where 'primitive' refers to this notion of irreducibility, rather than its meaning in evolutionary terms. For many philosophers, this is logically incoherent, and so agent causation is deemed to be conceptually and metaphysically impossible or at least highly problematic (Kim, 1993, 2000, 2006; see also Clarke, 1996). To see why, one needs to consider a number of different but overlapping ontological and methodological arguments that each appears to refute the concept from slightly different angles.

The first is the idea of *vertical reductionism* and the associated causal fundamentalism that typically comes with it. Vertical reductionism is the idea that for every macroscale description of a phenomenon, there is a microscale description that fixes it. In philosophy, this is often referred to as a supervenience relation: the macroscale supervenes on the microscale if a change in the microscale is necessary for a change in the macroscale to occur. Consider, for example, the relation between an individual molecule and the atoms that underlie or fix it. The properties of the molecule do not change without change at the atomic scale. If the microscale fixes the macroscale, then there is a sense in which the macroscale can always be derived from, or explained in terms of, the microscale. That is, any larger scale description of a phenomenon can be reduced to a description that uses smaller scales.

The logical endpoint of this vertical reductionism is that everything is describable in terms of the smallest possible physical elements in the universe—quantum fields, perhaps— and the interactions between them. This poses a threat to agent causation because the natural instinct is to infer that, because everything at a larger scale of description, including agents, is derivable from the properties of and interactions between quantum fields, what causal work is there left to do at the macroscale? Or as philosopher Jaegwon Kim (2006) puts it:

> "If an emergent, *M*, emerges from basal condition *P*, why can't *P* displace *M* as a cause of any putative effect of *M*? Why can't *P* do all the work in explaining why any alleged effect of *M* occurred?" (p.558).

If all causation is fixed by microphysical happenings, then agents must be epiphenomenal. There is simply no room in the universe for macroscale phenomena to have the kind of

'irreducible' causal power needed for agent causation (Kim, 1993; Bickle, 2015; Barwich, 2021).

The second challenge to agent causation comes from what we could *call horizontal reductionism*. Here, the threat to agent causation stems from claims that it is not the system, as a whole, that has causal power. Instead, causation is attributed to a specific subset of components *within* the system—it is a particular 'part' (or set of parts) that causally determines the action of the 'whole'. Thus, the whole does not have any causal power in virtue of being a whole because causation is entirely localised to (a set of) components with smaller spatial and temporal dimensions than the whole.

The important difference between vertical reductionism and horizontal reductionism, then, is that, in the former, the macro- vs. micro-distinction (*what* is being reduced to *what*) is between scales of description. However, crucially, at whatever scale you are using, the description is always of the *entire* system, that is, pitched at the spatiotemporal dimensions of the whole. Whereas, in horizontal reductionism, the macro- vs. micro-distinction refers to what is being described. The important causal elements in the system are reduced to a subsystem, or set of parts, that is spatiotemporally *smaller* than the system itself. Importantly, this can be carried out using any description scale. For instance, if the system in question is a brain, you could describe it in terms of all the atoms that constitute the brain and then causally reduce that to a localised subset of those atoms, or you could describe the brain in terms of its neuronal connectome and then causally reduce that to a localised subset of neural states.

The traditional alternative to agent causation, *event causation*, is one form of horizontal reductionism. It claims that an organism's actions are caused by states and events that simply *involve* the organism (Davidson, 1963; Franklin, 2018). Event-causalists tend to cite a particular psychological state (e.g., a belief) as the sufficient cause of a given action, with these states, themselves, having previously been caused by antecedent states or events. By horizontally reducing the locus of causality to events and states *inside* the organism, this view leaves the 'agent' out of the picture entirely. Agents are not identical with a particular belief or desire or goal, so if these psychological states wholly determine the system's behaviour, then there is no sense in which the 'agent' can be said to have caused its effects in the world. This is sometimes referred to as the 'problem of the disappearing agent' (Pereboom, 2004, 2014; Griffith, 2010; Franklin, 2014).

Eliminative materialism goes one step further than event causation, claiming that certain psychological states such as beliefs and desires do not exist and, thus, we should only appeal to neuroscientific explanations when identifying the cause of an organism's action

(Churchland, 1981; Ramsey et al., 1991). This position is, therefore, both vertically and horizontally reductive; in that it (vertically) reduces psychological states to neural states and (horizontally) reduces the whole system to a subset of neural states within the system. One issue worth noting is that it is not clear why eliminative materialism does not lead to the causal fundamentalism that typically comes with embracing vertical reductionism. In other words, what makes neural states the right scale of description at which to identify causation? If a reduction from psychological states to neural states is permissible, why not reduce further to a causal explanation at the scale of cellular components, electrical ions, or the subatomic particles and quantum fields that compose those neural states? As we have seen, once the door to vertical (causal) reductionism is opened, it is hard to see how one could resist the fundamentalist pull to only identify causation at the smallest scale of physical description.

The third and final threat to agent causation is from *external determinism*. If systems are determinately 'pushed around' by states and events in the environment, then to what extent can the system be meaningfully considered the cause of its actions? Instead, agents need to be causally autonomous from their environment, in that they have "the ability to do what one does independently, without being forced to do so by some outside power" (Boden, 2008, p.305). In this sense, external determinism parallels, and, in fact, overlaps with, the horizontal reductionist challenge to agent causation, since both threaten to 'explain away' the agent by localising causal power to somewhere that is *not* the system as a whole. In horizontal reductionism, it is events and states *within* the system that wholly determine the system's actions. In external determinism, it is events and states in the environment that wholly determine the system's actions.

These three lines of argument, vertical reductionism, horizontal reductionism, and external determinism, appear to underlie the general consensus among philosophers that agent causation is not tenable. Our primary goal in this paper is to demystify and revive the concept of agent causation by presenting a set of conditions that, in principle, would enable a theoretical system to overcome all three of those arguments if met. In other words, our aim is to propose a set of general criteria that may collectively justify ascribing agent causation to a system of study. These are:

1. Thermodynamic autonomy;
2. Persistence;
3. Endogenous activity;
4. Holistic integration;
5. Low-level indeterminacy;

6. Multiple realisability;
7. Historicity;
8. Agent-level normativity.

Our intention is not necessarily to argue that these criteria are complete and definitive but rather to demonstrate a plausible way in which agent causality can be conceptualised and realised in systems without violating the thesis of physicalism that generally underlies the scientific worldview. In doing so, the secondary aim of this paper is to investigate and argue that living systems satisfy at least some of these conditions. Note that we do not consider the criteria as bright lines but rather as dimensions along which different organisms may vary. Moreover, though we highlight each condition separately for the purposes of this paper, they do, in fact, overlap and co-depend to quite a considerable degree. While each is a necessary condition in its own right, the full sense in which the eight criteria may justify agent causation comes from understanding how they fit together as a collective.

We should also note that we take agent causation to be necessary but not sufficient for 'free will'. The criteria presented here are intended to provide a naturalised account of how a system may exhibit agent causation, but we do not assume that such a system would necessarily have free will (where the former may be loosely understood as 'doing what you want' or 'acting for your own reasons', while the latter may also include an ability to 'do otherwise' (O'Connor & Franklin, 2022) or a meta-capacity to 'want what you want' or 'reason about your own reasons' (Schopenhauer, 1839/1999)).

In short, we hope to (i) convince readers that agent causality is a plausible and appropriate way to think about causation in biological systems, and (ii) set out a conceptual framework for a more productive and empirically grounded investigation into the concept of agency within biology.

2.3. Criteria for Agent Causation

2.3.1. Thermodynamic Autonomy

One presumably uncontroversial condition for agent causality is that the system in question is thermodynamically distinct from its environment (known, technically, as being *out of equilibrium* with the environment). Indeed, this is necessary to be an organism or an entity at all—let alone be an agent or a cause of events in the world. Systems, first and foremost, must be identifiable *things*, with a physical boundary that legitimately separates

them from the rest of the universe (Schrödinger, 1944/2012; Barandiaran et al., 2009; Pross, 2016). As Keith Farnsworth (2018) explains:

> "If there were no physical boundary between an agent and its surroundings, then any external force would act unaltered throughout the agent, and any force initiated by the agent would act equally on the internal and external environment, so no distinction in action could be made between the agent and its surroundings. If no distinction can be made, then no measurement could be made to tell us whether an action arose from within or beyond the agent" (p.12).

Living systems achieve this separation from the environment via a cell membrane or an outer skin in more complex organisms. These physical barriers define the system as an entity, with a discernible inside and an outside, and, at the same time, enable the system to do work to remain out of thermodynamic equilibrium with the environment. The physical barrier also grants the organism a degree of causal insulation from the external milieu. Instead of being exposed to every flow of physical and chemical causation that comes its way, the system is sheltered from that storm and able to buffer its internal dynamics. This is a crucial first step toward agent causation because it means that events in the environment are not determinately pushing the system around. External forces may impinge on the system without leading to a change in its behaviour; thus, creating space for other (agent-level) factors to have a causal influence. In this sense, a physical barrier lays the groundwork for systems to overcome the threat posed by external determinism (see **Section 2.3.3**).

Moreover, the system's barrier opens up the possibility of a new form of causation, dependent on information. Rather than letting (potentially noxious) chemicals from the environment into the system, specialised receptor molecules can sit in (e.g.,) a cell's membrane. These receptors have both external and internal components, such that when the external end binds a chemical, it causes the entire molecule to change its conformation, which, in turn, can impact the dynamics *inside* the cell. No matter or energy passes through the barrier. Rather, the receptor transmits a *signal* that carries information about something out in the environment, namely, the presence of chemical *X*. The information is what causally impacts the system's internal processes and—as we explore below—this new form of *informational causation* can be heavily influenced by the current state of the cell (see **Sections 2.3.3** and **2.3.4**) and its historicity (see **Section 2.3.7**); such that the behaviour of the system is not wholly explainable in terms of environmental causes alone (*contra* external determinism), nor wholly understandable in terms of sub-localisable, instantaneous physical states (*contra* horizontal and vertical reductionism). Thus, the

requirement for agents to be thermodynamically autonomous systems sets the foundation upon which most, if not all, of the other agent causation criteria, are built.

2.3.2. Persistence

Another straightforward and foundational requirement for agenthood is that the system in question persists through time. While the duration necessary to satisfy this condition is an open (and probably unproductive) question, at least some degree of persistence is needed in order to be around long enough to cause an effect. Moreover, as we will discuss in **Sections 2.3.7** and **2.3.8**, persistence on phylogenetic and ontogenetic timescales are both required for systems to exhibit the type of *temporally extended* causation that, we argue, is needed to overcome the threat of eliminative materialism.

Living systems clearly satisfy the condition of persistence. Indeed, life is often characterised as the ability to (locally) resist the second law of thermodynamics and survive for extended periods of time (Schrödinger, 1944/2012). Unlike rocks which persist passively, due simply to physical hardness and chemical inertness (i.e., through stasis and inertia), biological persistence is dynamic in nature. Organisms obtain stability out of a constant internal flux of recursive chemical reactions and self-sustaining causal loops, where component parts do work to constrain one another within the bounds necessary to keep the whole system going. Living organisms thus persist as dynamic, holistic patterns that constantly regenerate the constraints required to keep themselves organised (Maturana & Varela, 1980; Juarrero, 1999, 2015; Montévil & Mossio, 2015).

There are two important consequences of the organism's method of persistence. First, the system needs the energy to fuel the self-sustaining cycles of constraints. Organisms must therefore remain open to the environment in a controlled way, locating food and converting it into energy that can be used to maintain the ongoing internal dynamics. Second, life is a pattern. What persists through time is the organisational pattern of self-maintaining, dynamical processes, *not* the physical material that realises that pattern at any one moment. In fact, the atoms that constitute the system are regularly being replaced, yet the dynamical pattern persists, and we take this to be the organism persisting.

This property immediately undermines horizontal reductionism since it is not clear how one could isolate causation to a subset of specific elements *within* the pattern when holistic integration is the system's defining characteristic (see **Section 2.3.4**).

2.3.3. Endogenous Activity

Non-equilibrium thermodynamic systems that persist through time can stand apart from the physical world long enough to be a causal factor in it. However, to justify agent causality and meaningfully be labelled the *cause* of their effects, systems also need to sidestep the challenge of external determinism. Their actions cannot simply be determined by events in the environment, such that one could predict the behaviour of the system from external states alone. In other words, agents need to have some degree of causal autonomy from the outside world.

As we saw in **Section 2.3.1**, a physical barrier is a good start: it insulates the system from external perturbations so that not every change in the environment causes a change in the system. However, it does not follow from this that every change in the system is not caused by a change in the environment. Thus, a physical barrier is not sufficient for causal autonomy since it is still the case that the forces that *do* causally impact the behaviour of the system could do so in a deterministic, linear fashion (as suggested by external determinism).

We propose that a condition for agent causality is not just persistence but active persistence. Systems need to internally initiate their actions, to some degree, so as to avoid being mere stimulus-response machines that are passively pushed around by the environment. In this way, external factors can *influence* but not determinately cause the system's behaviour in a manner that would be incompatible with agent-level causation (due to external determinism).

As we outlined in **Section 2.3.2**, a defining feature of living organisms is their dynamic, internal activity. They are constantly doing thermodynamic work internally to self-organise and maintain the pattern of processes that define their existence. To this end, organisms are better understood as actively monitoring and *adjusting* to information about conditions in the environment rather than being pushed around by these conditions. External inputs are assimilated into ongoing patterns of biochemical and/or neural activity rather than driving that activity itself.

Evidence of this sort of endogenous activity comes from all areas of biology. Central pattern generators (Lydic, 1989), for example, which *spontaneously* generate rhythmic oscillatory patterns in the brain, are found in almost all vertebrates and invertebrates and play an important causal role in organism-level behaviours such as walking, swimming, and flying (Guertin, 2012). In humans, our internal activity in the absence of external stimuli is abundantly clear whenever we introspect or sit alone with our thoughts. This has

been shown experimentally through reports of visual hallucination in sensory-deprivation contexts (Solomon et al., 1957; Flynn, 1962; Raz, 2013). The brain's constant activity is also evident in brain-imaging studies: fMRI research showed that task-specific neural activity typically amounts to just 1–2% of background brain activity in relevant areas and, thus, energy-consumption at rest is almost the same as during a demanding task (Raichle, 2010) (see Brembs, 2021 for a full overview of the brain as an endogenously active organ).

Taken together, this evidence presents a clear picture of organisms as endogenously active systems. Even when they appear from the outside to be at rest, they are not static internally (Dohmatob et al., 2020). They are constantly doing work to readjust their internal configuration so as to keep the whole dynamic, self-maintaining process alive and persisting. This equips the system with a degree of causal autonomy from the environment: external stimuli can influence the system's behaviour, but only within the context of its current internal dynamics. That is, the system's actions in the world are ultimately generated from *within*, with information from the environment used to guide these dynamics as needed (Buzsáki, 2019).

Another point to note here is that the information that organisms receive from their environments is very rarely sufficient for them to unambiguously determine an appropriate course of action. Often, external inputs are simply not rich enough for external determinism to hold. Instead, organisms actively probe the environment, gathering information, making inferences about what is out in the world, and trying out new problem-solving techniques. Thereby providing clear instances of endogenous action that is not determinately driven from the outside (Pezzulo & Nolfi, 2019). In this sense, even perception might be understood as a form of internally initiated action (Clark, 2015).

Therefore, living systems provide a model for conceptualising how systems, in general, can overcome the challenge posed by external determinism and, thus, move us one step closer to agent causality. In the rest of this paper, we will consider how a system might overcome the reductionist threat, starting with horizontal reductionism in the next section.

2.3.4. Holistic Integration

The horizontal reductionist challenge to agent causation paints a picture of systems as machine-like, made up of isolatable mechanisms, each of which performs its own specialised function, which are then combined in a serial, linear, additive fashion to create the system (or *cause* it to come into being). This sort of explanation is "analogous to a recipe for producing a phenomenon starting from a list of ingredients, where the ingredients are mechanistic entities and their properties, and the recipe amounts to the

organization and sequence of activities these entities perform" (Băetu, 2015, p.105). This perspective leads to the 'agent' being left out of any causal explanation for its own behaviour because, for any given action, there is an identifiable mechanism, or linear cause-effect pathway, *within* the system that can be pointed to as the cause of that action. Thus, while the system may have escaped being deterministically pushed around by external events, it is now being deterministically pushed around by some of its own component parts instead. There is no room left for a causal power that inheres at the level of the whole system (i.e., agent-level causation).

However, we contend that not *all* systems can be coherently decomposed into separable parts in this machine-like manner. As we saw above, living organisms are dynamic, holistically integrated systems whose parts constantly act in concert, influencing and constraining one another in order to maintain the holistic pattern. Even the simplest organisms show substantial degrees of integration. For example, bacterial chemotaxis, when a bacterium locomotes up or down a chemical gradient in its environment, is one of the simplest and most well-studied behaviours in biology. There exists a well-understood pathway in the bacterium that links a transmembrane receptor for detecting food substances, via an internal signal transduction cascade, to a flagellum that controls its motion (Falke et al., 1997). The temptation is thus to explain bacterial chemotaxis in terms of a simple linear chain of fully determining causes and effects, starting with the detection of a chemical stimulus and then sequentially moving through each part of the chemotactic pathway until a locomotion behaviour is performed. However, the apparent discreteness and linearity of this pathway, as suggested by highly controlled (and thus, highly artificial) experiments, is somewhat illusory. In fact, contextual information about the current metabolic state of the cell constantly integrates with and modulates the chemotactic pathway (Porter et al., 2011) in a way that is highly nonlinear and that challenges any suggestion that the pathway determinately causes chemotaxis. If changes in the metabolic state of the cell can change how the action is executed, despite the same individual stimulus being present, then it does not seem accurate to isolate the chemotactic pathway and identify it as the cause of the action. Such an interpretation only arises when one holds the metabolic state of the cell constant while studying how bacterial chemotaxis works. As we have seen, this would afford only a partial understanding of the mechanism because it misses the system's holistically integrated and relational structure, as well as the endogenous activity of the system.

Bacteria can also integrate a number of different environmental cues simultaneously, including pH levels, temperature, and osmolarity, and use information from the past to

inform and influence action (Sourjik & Wingreen, 2012). It was even shown in *E. coli* that crosstalk between one signalling pathway (heat shock) could modulate the activity of a significantly different pathway (respiration) (Tagkopoulos et al., 2008; Lyon, 2015). Moreover, the direction of bacterial chemotaxis itself relies on temporal integration. The bacterium does not act based on absolute levels, but according to information it gathers about concentration changes as it moves. The resultant picture is one in which, even in this very simple organism, "sensory, regulatory, and metabolic networks must all drive environmental perception and corresponding action" in a strikingly holistic and integrated fashion (Freddolino & Tavazoie, 2012, p.364). The argument that the bacterium's behaviour is determined by any identifiable 'part' or even an isolatable pathway is thus difficult to maintain when the whole organism is clearly involved in the regulation and execution of this behaviour.

We see increasing degrees of holistic design as we scale up to more complex organisms with brains and nervous systems. Fuelled by recent technological advances that allow researchers to record neural activity on a much larger scale than ever before, we are now learning that brain areas we once understood as single-mindedly carrying out their work in relative isolation are, in fact, highly sensitive to the activity in other brain areas. Visual cortex activity, for example, as well as processing visual information, is substantially modulated by an organism's movements; with evidence coming from running and hindlimb flexions, all the way down to small orofacial movements and pupil dilations (Erisken et al., 2014; Musall et al., 2019; Urai et al., 2022). Indeed, information about actions, goals, diverse sensory percepts, internal states, and other parameters is much more widely distributed than previously thought, suggesting that all local signalling is likely being modulated by the context of global brain states (Steinmetz et al., 2019; Kaplan & Zimmer, 2020).

Similarly, the brain regions, circuits, and processes that mediate decision making and action selection are highly distributed, involving ongoing signalling between multiple subsystems across the brain, working in parallel in recursive, interlocking cycles over some duration of time (Grillner & Robertson, 2016; Silva & McNaughton, 2019; Steinmetz et al., 2019; Cisek, 2022). The result is a dynamic, system-wide interaction, in which parts are continuously modulating and constraining each other until they all *collectively* settle into a new state. To decompose this process into a set of functionally independent (machine-like) 'parts' carrying out their work in isolation, and then combining to cause the system's next state, is to entirely miss the essentially dynamic and holistic nature of it.

Therefore, the evidence suggests that biological systems are too holistic, too integrated, and too relational to submit to a machine-like analysis. Instead, they are more akin to Stuart Kauffman's concept of a 'Kantian Whole' (which, as the name suggests, derives from the work of Immanuel Kant in his *Critique of Judgement*). Kauffman (2013) depicts a system in which "the parts exist for and by means of the whole, and the whole exists for and by means of the parts" (p.609). That is, a system so deeply interconnected that it does not make sense to decompose it into its component parts because the true essence of the system exists in the relations between those parts (Dupré, 2011). In this sense, organisms offer us a way to conceptualise a system that does not fall prey to the horizontal reductionist challenge. If a system is so holistically integrated that to understand any given part, you must also understand the whole, then causal power will meaningfully inhere at the level of the whole. You cannot horizontally reduce such a system to identify a particular part (or set of parts) that is determining the system's next state because the activity of that part is, itself, determined by all the other parts in the whole. Therefore, we suggest that holistic integration is a necessary condition for agent causality.

2.3.5. Low-Level Indeterminacy

If the arguments above deflate *horizontal* reductionist claims, it could still perhaps be argued that the 'real' causation inheres at the lowest level, in the physical interactions of the components, even if they have to be considered as a single, unified dynamical system (no matter how complex those interactions might be). Under this *vertically* reductionist view, the state of the entire physical system at time t (which may need to include the state of its environment), plus the low-level laws of physics, fully determine the state of the system at some subsequent time, $t + 1$. Logically, this would extend to $t + 2$, $t + 3 \ldots t + n$, and on to infinity. This Laplacean view (of complete determinism from the dawn to the end of time) is encompassed in Kim's argument that 'every physical effect has a sufficient physical cause', with the implication that those causes are necessarily located at the lowest levels (or smallest scales of description) (Kim, 1993; 1998). This argument thus asserts both physical pre-determinism and causal reductionism. If the laws determining the interactions of particles or the evolution of quantum fields are causally comprehensive, then no higher-order, macroscopic causes can be admitted.

However, the evolution of quantum fields is *not* fully pre-determined in this way. Though the Schrödinger equation gives a definite solution for how quantum fields comprising any system will evolve, this solution is a distribution of *probabilities* (Rovelli, 2021). As soon as some interaction occurs of the type necessary to actually observe the state of the system (but not actually relying on an observer, *per se*), some particular set of these probabilities

will be realised. What determines what set becomes realised (within the probability parameters as defined) seems to be truly random—unpredictable *in principle*, not just in practice, consistent with the Heisenberg Uncertainty (or Indeterminacy) Principle. Though there are very different interpretations of what this means for the underlying nature of reality (Rovelli, 2021), the upshot is that the totally defined microscopic state of a complete system, together with the fundamental equations that comprehensively describe the evolution of quantum fields, as an empirical fact, *do not* deterministically predict the next state of the system (Del Santo & Gisin, 2019; Smolin & Verde, 2021; see also **Chapter 4**). This thereby brings into doubt the causal fundamentalist urge to identify all causal power at the system's lowest level of description.

This fundamental indeterminacy is, therefore, a necessary but not sufficient condition for the emergence of some kind of higher-order, macroscopic causation (Ellis, 2008; Steward, 2012; Mitchell, 2018b). Many physical systems will evolve simply according to the partly random realisation of these underlying probabilities. However, others may come to be structured in such a way that constrains their evolution according to some kind of higher-order, functional criteria. This is exactly what happens in living organisms, due, in the first instance, to the iterative, ratchet-like action of natural selection and, second, to the ongoing learning that individual organisms engage in over their lifetimes (Mitchell, 2023a). Both of these processes select for structures that physically embody criteria for action (see **Section 2.3.6**), enabling organisms to accumulate causal power and act in ways that promote their own persistence (see **Section 2.3.7**).

As we will see in **Section 2.3.8**, this goal-directedness of living organisms means that higher-order states can have meaning and value (normativity) for the organism. The system can be configured in such a way that it represents and operates on information—physical states that are *about something* and which inform the actions of the agent. This kind of informational causation thus enables agents to do things for reasons. However, if this view is to be defended, the causation must depend on the meaning of higher-order patterns, which, though it must be realised in some low-level state at any moment, cannot be reduced to low-level details. Here, the concept of multiple realisability is crucial.

2.3.6. Multiple Realisability

A key feature of agency is doing things *for reasons*. The lowest-level details of a system, e.g., the neurons in a brain or the atoms and molecules that constitute them, do not have reasons; certainly not *reasons for* system-level action, at least. Therefore, a system cannot justify agent causality if it is being entirely driven around by the constituents at its lowest level of description. Instead, higher levels need to have some degree of causal autonomy

from the lowest-level parts, such that changes at the lowest level do not *necessarily* cause changes at the higher level in a simple feedforward or bottom-up fashion. This is a crucial step toward agent causality because it means that at least some of the system's causal power must inhere at the higher levels of organisation, where it might start to make sense to talk about agent-level *reasons*.

The structuring of criteria in living organisms, which determine the flow of information and interpretation of different patterns of activity, embodies exactly this kind of higher-order causation (Tse, 2013). This is particularly elaborated in neural signalling, where communication from one neuron to another, or one population of neurons to another, depends on distinct macroscopic *patterns*, which each can be instantiated or realised in many different microscopic arrangements (i.e., multiple realisability).

Take the communication between neurons, for example. Neuron A receives inputs from Neuron B via neurotransmitter molecules that are released when B fires. These molecules can be bound by neurotransmitter receptors on Neuron A, which open, allowing sodium ions to rush into the post-synaptic structure. This generates a change in the electrical potential gradient between the inside and the outside of the cell. However, this charge is rapidly buffered unless the total increase over some short time period pushes the potential above a threshold, which triggers an explosive amplifying event—the action potential or firing of the neuron, which then transmits an electrical signal to the other neurons it has synaptic connections with (also known as a spike). If the threshold is *not* reached, however, the neuron will not fire, and the electrical potential will peter out, in effect wiping any record of neuron B's original spike.

Neuron communication is, therefore, generally not designed to be a faithful or exact transmission of 'spiking' information but rather an interpretation of it. Some spikes will be completely ignored and for those that are not, details such as the timing of a spike are still often ignored, i.e., the specific details of exactly *when* the spike occurred simply do not matter (outside of some specialised systems such as auditory neurons). Instead, neuron A is only interested in whether the number of spikes that come its way within a given time frame is sufficient to satisfy its threshold—or, as neuroscientist Peter Tse (2013) frames it, sufficient to meet the *criteria* neuron A places on its inputs. The process is thus highly nonlinear; in particular, many changes in the input do not lead to changes in the output. Individual neurons are configured so that they monitor their inputs in different ways; some act as temporal filters, waiting for strong bursts of inputs over a limited time window before firing, others act as coincidence detectors, where near-synchronous spikes from multiple input neurons are needed to meet their conditions for action. The low-level

details thus often do not matter, and they are often lost in this coarse-graining process. As a consequence, two patterns of input spikes with different arrangements may effectively prompt the same activity in the downstream neuron. In effect, they *mean* the same thing to that neuron, and it is the meaning of the pattern that has causal efficacy.

The same principle applies at the level of populations of neurons (Saxena & Cunningham, 2019; Ebitz & Hayden, 2021). Because of the network of excitatory and inhibitory connections among any local population of neurons, the possible patterns of activity that groups of neurons can exhibit is constrained (Deco & Rolls, 2006; Semedo et al., 2020; Ebitz & Hayden, 2021). Neuronal populations within a brain region consequently tend to occupy a particular set of patterns, or 'attractor states', far more often than any of the other patterns they could conceivably occupy. This means that the potentially high-dimensional state space that the population could occupy is constrained to a much lower-dimensional manifold. Brains are shown to capitalise on the occurrence of these population-level attractor states to, once again, use higher-order information to drive the system, rather than rely on the low-level details (in this case, the states of the individual neurons). A given population can exhibit a number of neural modes or subspaces of the possible multidimensional activity space, only a subset of which may be efficacious in driving activity in downstream areas (Semedo et al., 2019). Perhaps a more helpful way to frame this relationship is that the second population of neurons is configured to monitor activity in the first region and selectively activate when certain patterns appear while effectively ignoring others (Buzsáki, 2010). A different downstream population may only activate when it detects a different set of patterns, depending on what information it is "interested in".

These functional criteria can be prewired by evolution or crafted through learning. In either case, they are aligned to the goals of the organism, better enabling the organism to do the causal work required for its own persistence. Moreover, they can also be modulated on the fly by processes of attention or top-down selection, realised through rapid synaptic reweighting (Tse, 2013). The functional criteria that direct each neuron's or each population's interpretation of any given inputs are thus not fixed or isolatable but depend holistically on global context (thus, also *contra* horizontal reductionism).

Multiple realisability of this sort undermines claims that the causality really rests in the details of the microphysical goings-on. The vertically reductive argument is that even if some low-level pattern becomes coarse-grained in transfer to the next neuron or population, since the macrostate must, at any moment, or over some defined duration, be instantiated in *some* particular microstate or sequence of microstates (i.e., it supervenes

on the microstates), then any particular microstate will necessarily *fix* the macrostate. Therefore, the causality that we had attributed to the macrostate A (under the relationship: if A, then X) could really be completely inherent in (and, thus, reducible to) the corresponding microstate, A′, which would also entail X. However, there is a reciprocal relationship that is usually implied in this kind of causal relationship. If A is supposed to be *the cause* of X, then it is usually understood that the counterfactual NOT A would imply NOT X.

Crucially, under multiple realisability, that counterfactual applies to the macrostate but not necessarily to every microstate. If the microstate is changed from A′ to B′, corresponding to macrostate B, rather than A, then you may obtain a different outcome, Y. However, if the microstate is changed from A′ to A″ or A‴, all of which still corresponds to macrostate A, then you would *not* see a change in the outcome—X would still occur. The effect, X, is therefore 'sensitive' to the multiply realisable macrostate, A, rather than the microstate that physically realises it (i.e., A′, A″ or A‴) (Sinnott-Armstrong, 2019). The causal information, under this kind of counterfactual reasoning, is thus rightly said to inhere at the level of the macrostate, i.e., in the higher-order pattern, rather than the low-level details.

Taken together, low-level indeterminacy and multiple realisability appear to bypass the vertical (causal) reductionist challenge to agent causality. Low-level indeterminacy shows that the lowest level of description is not deterministically fixing the next state of *any* system, while multiple realisability represents a way in which *some* systems can capitalise on this indeterminacy in order to exert a form of higher-order, macroscopic causation. In the sections below, we argue that these multiply realisable, higher-order patterns have causal efficacy in the brain by virtue of what they *mean* for the organism. This provides a framework for naturalising *reasons* in systems more generally.

2.3.7. Historicity

Above, we described how low-level indeterminacy creates some causal slack in the system that can enable the emergence of non-reductive (or higher-order) causation. But does this actually get us any closer to agent causation? Even though it is higher-order patterns of information that push the system around and not the lowest-level physical happenings, it is still not obvious how the causal power in the system could sensibly be understood in terms of *agent-level reasons*. Is the whole not still being 'pushed around' by its parts? It is just being carried out by *patterns* of neural activity now. In this case, there is still no sense that the 'agent' is causing its effects in the world.

To bring the agent into the picture, we need to reflect on our framework for understanding causation in contemporary science. Mostly, we limit our causal explanations to 'how' questions: we ask '*how* did X cause Y?' Scientists then try to get to grips with the mechanisms that underwrite X's behaviour, and we explain the causal relation between X and Y in those terms. However, an equally valid question to ask is: '*why* does X behave in that way?' (Dretske, 1988). In the case of rocks or atoms or billiard balls, this is a somewhat uninteresting and unenlightening question, with the answer simply being the physical characteristics of the entity in question. However, in systems that run on information, it is a very fruitful line of questioning, precisely *because* these systems exhibit higher-order, macroscopic causation.

Consider, again, the neural population. An activation pattern among the neurons in this population is only causally efficacious (within the system) if there is a receiver or interpreter monitoring the population and activating in some way when the pattern appears (Buzsáki, 2010). That is, the pattern is only informational if there is something that is sufficiently configured to notice it. Therefore, it is relevant to ask *why* the interpreter is set up to detect the particular pattern it does (and subsequently activate in the way that it does) if we are to obtain a full causal understanding of the system. The answer, we suggest, is because of what the pattern *means*, not in any kind of immaterial or mystical sense, but in a way that is grounded in the system's interactions with the world, shaped by its history, and instantiated in its physical structures.

In biology, natural selection infuses organisms with the purpose to persist. If organism A is set up such that it out-persists organism B, then future populations are going to look more like A than B. Over time, the set of surviving organisms is going to be the one that is best set up to persist; that is, the set of organisms whose value system, what is seen as 'good' and what is seen as 'bad', is most optimally conducive to each organism's persistence, given its historical environment (compared with all the other possible ways in which that value system could be set up). In other words, a lineage's successful persistence over evolutionary timescales cashes out as *meaning* or normativity in the individual. For example, the food chemical *means* something good to the bacterium because it was adaptive in the past to move toward it, and so it is set up to do so again. The noxious chemical *means* something bad because it similarly was adaptive to locomote away from it in the past, and so it is disposed to do so. This meaning is grounded in its ancestors' previous interactions with the world and defined relative to their persistence; such that what the stimulus means to an organism is often well aligned with the effect that stimulus

is going to have on its chances of survival (based on the *actual* effect it had on previous generations).

In simple systems, the meaning of a signal can be pragmatically embodied through direct coupling to some historically adaptive action. The first nervous systems, presumably similar to the simple nerve nets observed in extant creatures such as Hydra and jellyfish, may have co-evolved with muscles to allow coordinated movement of the various parts of multicellular creatures (Mitchell, 2023a). Linking these to sensory receptors could then allow appropriate responses to be selected based on information about the external world in the context of the animal's own movements (Keijzer et al., 2013; Dupre & Yuste, 2017; Jékely et al., 2021). As organisms increased in complexity, perception and action became decoupled, with the addition of intervening layers now operating on internalised 'representations'. These internal, representational patterns are still grounded by links to the periphery in both directions (from perception and to action); that is, they stand in exploitable relation to things in the world and adaptively inform action but they can now be integrated and operated on in more complex ways (Millikan, 1984, 1995; Shea, 2018).

Returning to the neural population, we now see that the interpreter is set up to recognise and act in response to a given pattern *because* it is meaningful. If the pattern co-varies with a state of affairs in the world that can be used to enhance survival, then being configured to detect that pattern and use it to guide appropriate action in the world is adaptive. In many cases, the particular neural pattern that comes to represent or "be attached to" some referent is arbitrary, selected from a set of preconfigured endogenous neural ensembles that happen to be active at the time of some experience (Buzsáki, 2019). The semiotic relationships in the nervous system are thus not driven from the outside or determined by physical properties (Joslyn, 2000). Yet, treating the pattern as meaningful and turning it into information that causally influences the next state of the system is going to lead to increased persistence, and thus, over time, lead to a higher percentage of the population being configured, not only to treat the pattern as meaningful, but for it to mean something relevant to action and survival.

In effect, for systems that run on informational causation and emerge from an evolution-like selection process that rewards persistence, it is *meaning* that drives the system and, thus, informs behaviour. Moreover, the meaning inheres not just in the higher-order pattern itself but also in the recognition and response to that pattern. This view therefore undercuts vertically reductive, eliminative claims that behaviour can be reduced to the flow of patterns of neural activity, with the mental content or meaning of the encoded states being effectively epiphenomenal. In fact, it inverts that logic. Neural patterns have

causal power in the system *solely by virtue of what they mean* to the organism as a whole (see **Chapter 3**).

Any attempt to understand or explain the causes of an organism's behaviour is thus doomed to fail if it takes a purely instantaneous view of the physical system. It is not enough to account for how an organism behaves upon detecting some external stimulus or physiological state of affairs—the 'triggering cause'. We must also understand why the system is configured such that it behaves in that way—the 'structuring causes' (Dretske, 1988). The actual causal influences are diachronic; that is, they extend through time. What gives organisms causal power is the evaluative record of their past experiences and those of all their ancestors (as well as, in a negative sense, all those unfortunate individuals who did not leave offspring). This causal power is hard-earned. Both natural selection and learning do causal design work, they configure the system in such a way that it embodies pragmatic knowledge about the world and itself that can be used to direct adaptive action. Living organisms thereby accrete causal power and come to act as causal agents in the world.

Crucially, agents do not just learn how to respond to things in the world and then sit waiting for those stimuli. They adapt their endogenously generated patterns of active behaviour to their environment and circumstances, learning what to learn from, what information to seek out, and what active tactics and strategies to use to best pursue their goals, crafting their own environments as they do so (Cisek, 2019).

If we consider nervous systems as control systems, then we can broadly characterise the meaning of various neural states as representing beliefs (about things in the world or the state of the organism itself), desires (usually with short-term goals nested within a framework of longer-term goals), and intentions (possible actions that may be considered, evaluated, and ultimately selected for execution or not). Together with the foregoing discussion, this seems to yield a naturalised account of how a system can do something *for its own reasons*. Note that these reasons do not have to be conscious in order to justify agent causation (and mostly, they will not be). We contend that to be a system in which *meaning* is the causal driving force just *is* what it is to have reasons in the manner necessary for agenthood.

However, there is a remaining challenge from some theories of event causation.

2.3.8. Agent-Level Normativity

In complex organisms with nervous systems, what determines the flow of brain activity from state to state is the *meaning* of the neural patterns. The idea that what an organism

does can be reduced to the neural patterns themselves is therefore not tenable; the system is configured such that those patterns only have causal power *by virtue of what they mean*, i.e., the psychological states that they correspond to (conscious or subconscious). However, even if that inability to reduce decision making to neural states is accepted, some proponents of event causation argue that *psychological states*—beliefs, desires, and intentions—can do the necessary causal work. Under this view, the collection of such states at any given moment constitutes the 'events' that determine what happens next; the agent is simply the *arena* in which such states arise (cf. Ekstrom, 1993; Kane, 1999; Franklin, 2014, 2018).

We argue here that this is not the right view to take. Beliefs, desires, and intentions are things that only an agent can have. Neurons do not have beliefs; neural circuits do not have desires; brains do not have intentions. Moreover, an intention is not a thing (either a substance or an event) that can either exist or have causal power by itself. It is *the agent having the intention* that has causal relevance to what happens. When the entire set of such psychological states is taken into account at any moment, holistically integrated, as described in **Section 2.3.4**, and freighted with historically grounded meaning as described in **Section 2.3.7**, we contend that this *just is agent causation*.

The agent *itself* is the locus of meaning. That is, meaning inheres at the level of the whole system, as an entity persisting through time, interacting with its environment and being judged on its behaviour. Percepts, drives, actions, and their consequences mean something for the whole agent, not for its parts. This derives from the fact that, from an evolutionary perspective, the whole organism is the locus of fitness, not its parts. Natural selection thus crafts the control systems of living organisms to detect, characterise, and operate on signals *as they are relevant for* the survival of the whole organism. The chemotactic system of *E. coli*, for example, is configured in the way that it is because it has been adaptive for millions of preceding generations to move up a concentration gradient of a food source. The value and meaning in that relationship are grounded relative to the purpose of the entire system, which is to persist as a unified whole.

In the transition from unicellular to multicellular life, the locus of fitness shifted from single cells to the multicellular whole. This transition involved a progressive division of labour within clonal organisms, first in transiently aggregating colonies and then in obligate multicellular creatures. In particular, the division of soma and germline means that individual somatic cells give up any chance of reproducing directly. However, their genetic material can be reproduced indirectly because it is shared with the germline cells. Natural selection thus ceases to care about the fate of individual cells in the multicellular

organism; all that matters is that the organism as a whole survives and reproduces (Michod, 2007; Folse, 2010), and the functional roles of all cells are directed towards that end.

In animals with nervous systems, cumulative feedback from natural selection configures the nervous system to enable organisms to detect stimuli and action possibilities in the environment that are most salient for the survival and reproduction of the whole organism. Perception is egocentric, action-oriented, and laden with value from the get-go. The goal is to create a map of objects out in the world that represent potential threats and opportunities *for the whole organism*. In parallel, the action selection systems are fundamentally configured around actions of the whole organism—approach, avoidance, exploitation, exploration—and whether those actions tend to be good or bad *for the whole organism*.

On both evolutionary and individual timescales, the agent is thus the locus of fitness. It thereby becomes both the locus of meaning and also the locus of control that is informed by and evaluated relative to that meaning. Of course, natural selection has not just given animals hard-wired instincts. It has endowed them with systems to *learn* from their experience, in particular systems of reinforcement learning that will up- or down-weight action choices based on the outcomes of prior behaviour. Again, the value of these outcomes, the thing that grounds the meaning of possible actions, is relative to and inheres at the level of the whole organism. Individual neurons do not feel rewards or punishments; the agent does, and it is the agent that is guided by these signals.

2.4. Summary

To summarise, we argued that the eight criteria outlined above constitute a completely naturalistic way for systems to, in theory, exhibit agent causation. To recap, systems can be agents if they are self-organising and causally insulated enough to persist through time, out of thermodynamic equilibrium with the environment. To avoid external determinism, they need to be intrinsically active, treating external inputs more as helpful information than determinate, causal forces. The proactive self-organising activity of these systems entails a holistically integrated structure, in which parts are too interconnected and context-dependent to be understood in a machine-like, decomposable, linear fashion (*contra* horizontal reductionism). On top of this, these systems can be driven by meaning and reasons because higher-order organisational patterns are able to coarse-grain over microphysical happenings by virtue of the existence of some degree of indeterminacy at

lower levels. This meaning derives from temporally extended causal processes that shape the physical structures of the system to reflect and respond to information about the world that is relevant to it. The meaning of higher-order states is thus grounded in historical interaction with the environment and attuned to given selection criteria (e.g., in natural selection, the criteria is persistence and reproduction). We contend that a system whose activity is informed by this kind of higher-order meaning just *is* a system that is exhibiting macroscopic causation and acting for reasons (*contra* vertical reductionism). Additionally, these reasons are rightfully understood as inhering at the level of the whole system because that is where the locus of fitness is that selects for them, that level is what those reasons are about, and so that is where the appropriate locus of meaning and causality lies (*contra* event causation).

This set of conditions, if collectively met, avoid all three of the arguments set out in the introduction (**Section 2.2**)—external determinism, vertical (causal) reductionism, and horizontal reductionism (including event-based causation). Moreover, it does so in a perfectly naturalistic way, without the supposed mysticism or dualism that accounts of agent causation have often been accused of. Perhaps most notably, it meets the challenge of Kim by showing that the causation in the system is *not* wholly inherent in or captured by the low-level details of all the physical components at any given moment. As living organisms demonstrate, any such picture of the instantaneous state of an agential system misses the extended history of causal influences that imbue the states with meaning relative to the goals of the whole system, meaning on which the agent selects its actions. Thus, it would be wrong to reduce causation to the system's lowest level of description.

In setting out these criteria for the justification of agent causation *in principle*, we simultaneously argued that biological organisms, *in practice*, may satisfy most, if not all, of them. To reiterate, while we used living systems as model examples to help conceptualise how each condition could conceivably be met, we did not intend to imply that living organisms are necessarily the only things that could qualify as agents. The primary purpose of the paper was to set out a plausible way in which agent causation could be realised or naturalised in *any* theoretical system and in doing so, lay the groundwork for future research that uses the framework to evaluate the agency of individual organisms or systems.

Indeed, we take the criteria presented here to be dimensional, as opposed to categorical. We would therefore expect that agents vary considerably in how they satisfy each condition, with different systems perhaps performing strongly on some criteria but not on others. In this sense, existing mathematical methods and formalisms for measuring agency,

and other related parameters such as autonomy, integration and consciousness, are well suited to quantifying the sort of multidimensional agency we have outlined here. Different information-theoretic formalisms of *causal emergence*, for example, may be applied to measuring higher-order, macroscopic causation within a system (Hoel et al., 2013; Hoel, 2017; Rosas et al., 2020; Klein & Hoel, 2020). Similarly, attempts to formalise notions of individuality (Folse, 2010; Krakauer et al., 2020) and autonomy (Seth, 2010; Albantakis, 2021) can be used to measure *how* endogenously active and free from determinate external forces a system is, by quantifying the degree to which it may be more "interested in itself rather than the world outside" (Buzsáki, 2020, p.3). Finally, the criteria of holistic integration and informational causation more generally fits neatly with the long-standing measures associated with Integrated Information Theory (IIT) (Oizumi et al., 2014; Tononi, 2015; Tononi et al., 2016).

In sum, then, the framework presented here should help researchers working on measuring agency to situate their findings within the context of agent causation. If the analysis offered above is accurate, then agency is a multidimensional, multiply realisable concept that cannot be quantified in terms of a single parameter such as thermodynamic autonomy, system-wide integration or emergent information. It is a composite concept, with systems likely exhibiting different 'agency profiles' to one another and possibly even differences across an individual's lifespan.

2.5. Addendum: The Reductive Instinct

The idea that organisms can be causal agents, that they can act in the world, is entirely in keeping with common thinking. Even very young children naturally attribute agency to objects that appear to be moving under their own power and acting in an intentional manner (Keleman, 1999; Gergely & Csibra, 2003). The framework we outline above provides a naturalistic way to think about how such causal agency could emerge over phylogeny and ontogeny. Moreover, this view of agents is invaluable and arguably indispensable in building explanatory theories of organismal behaviour (Dennett, 1987; Steward, 2012; List & Rabinowicz, 2014). Why then is this view seen as problematic in philosophical circles, and why do so many scientists tend to fall into a more reductive, mechanistic way of thinking that seems to eliminate agents from the picture?

Part of the reason may be a slippage from methodological into theoretical reductionism. Biologists investigating the processes of life naturally try to isolate single processes from the ongoing dynamics of the cell or organism and, further, try to isolate individual

components of those processes to understand their functional roles and, ultimately, the logic of the entire process. To this end, they employ experimental techniques that powerfully manipulate individual components and measure some specific outcomes while attempting to hold as much of the background activity of the system constant. This horizontally reductive approach is (apparently at least) extremely powerful, for example, in delineating biochemical pathways or neural circuits mediating diverse cellular or organismal functions and in assigning roles to the many components of such subsystems.

This approach naturally lends itself to thinking that an organism's components truly act in isolation from each other. Even in philosophy, there is a tendency to build logical propositions, normative frameworks, or tightly constrained thought experiments that consider properties or events in isolation when trying to understand the causal logic of a system or tease out individual causal determinants in absolutist terms (Wimsatt, 2007).

However, just because it is possible and often useful to experimentally or conceptually isolate components or pathways or processes or properties, and to consider them separately, while holding everything else constant, does not mean that these elements actually "work" separately or have truly isolatable causal efficacy in the normal course of things. Even the use of a term such as "working" may give an overly mechanistic framing (Nicholson, 2019). Reductive approaches foster an illusion of linear pathways with dedicated components. However, any such picture relies on a forced perspective. Adopting different experimental or conceptual perspectives usually reveals that the cellular components or neural circuits are functionally involved in many different functions, that they have more promiscuous interactions than revealed from a single angle, and that both their activity and the consequences of their activity are highly context-dependent, integrated with the activity of other components and subsystems.

The challenges in translating basic research findings obtained using these kinds of reductive approaches to the clinic (Wong et al., 2019), in areas from cancer to psychiatry to neurodegenerative disorders, highlight the degree of hubris in thinking that complex dynamical systems can truly be decomposed and that manipulations performed under controlled conditions in the lab will have equally predictable and controllable outcomes "in the wild". The apparent successes of the reductive methodological approach thus need not, and we argue, *should not* entail a commitment to theoretical reductionism.

Chapter 3

Beyond Mechanism: Extending Our Concepts of Causation in Neuroscience

---

Status

This chapter has previously been published as:

> Potter, H. D., & Mitchell, K. J. (2025), Beyond Mechanism—Extending Our Concepts of Causation in Neuroscience. *European Journal of Neuroscience*, 61: e70064. https://doi.org/10.1111/ejn.70064

Author contributions

Equal contribution—both authors conceived, wrote, and edited the manuscript together.

3.1. 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.

3.2. 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 that at the lower scales of molecular and cellular interactions or the higher scales of neural circuitry, brain regions and macro-scale networks" (Ross & 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 support this program of mechanistic localisation and decomposition (Silberstein & Chemero, 2013; Silberstein, 2021) by providing complementary evidence that shows 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.

In 2005, the search for the neural causes of behaviour received its major boost with the invention of optogenetic technologies (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 (Woodward, 2005; Pearl, 2009), 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. Siemian et al., 2021; Filipowicz et al., 2022; Castaneda et al., 2024; but cf. Gomez-Marin, 2017; Yoshihara & Yoshihara, 2018). 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).

The important 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 behaviour 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* an animal's 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 understand 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., 2017, p.1). Or several recent studies that have elegantly applied systematic optogenetic activation of individual neurons or sets of neurons across the nematode and fruit fly nervous systems in order to derive the map of responses that follow from each such activation. In the nematode case, 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, our emphasis).

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 Dewey, 1896; Cisek, 1999; Buzsáki, 2019; Cobb, 2020; Brembs, 2021). 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.[9] 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,

[9] 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'.

physiologically considered, but a machine for converting stimuli into reactions" (James, 1890, p.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.

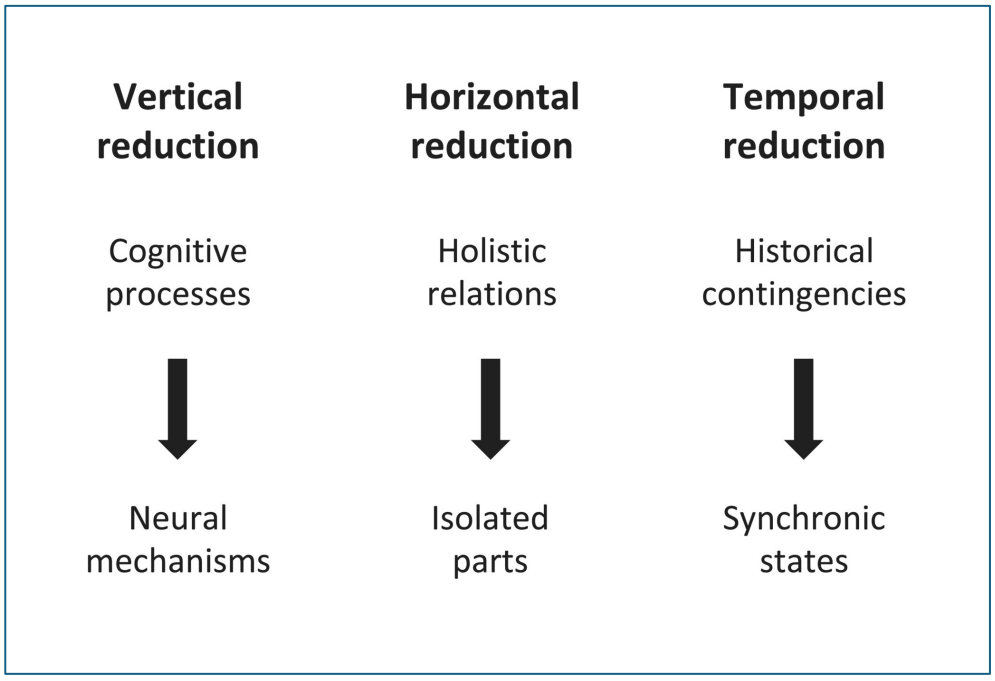

Figure 1 | Varieties of causal reduction. Taking neural mechanisms to be the explanatorily relevant causes of a behaviour entails a vertical reduction in ontological levels, from the cognitive to the neural, a horizontal reduction, involving isolation and decomposition, and a temporal reduction, with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

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*, thereby allowing us to effectively ignore the 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; **Chapter 2,** published as Potter & 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, 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 behaviour?'. 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.

3.3. 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 & 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 & 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 therefore 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. And most of the occurrent dependence conditions, such as the tensile strength of a window, are generic, inherent, and familiar properties of the system. Hence, they lack sufficient 'causal prominence' (Tseng & 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—like the tensile strength of a window. 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 subjective meaning.

3.4. 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 (2013), this is not a complete picture of how neuronal communication works. 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*.[10]

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 & 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

[10] 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.

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 we are describing is in fact entirely compatible with a driving view of causation after all.[11] 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**)**.** In other words, the nature of neurophysiology forces us to consider how and why a neuron comes to be configured such that it responds to its inputs in the way that it does.

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

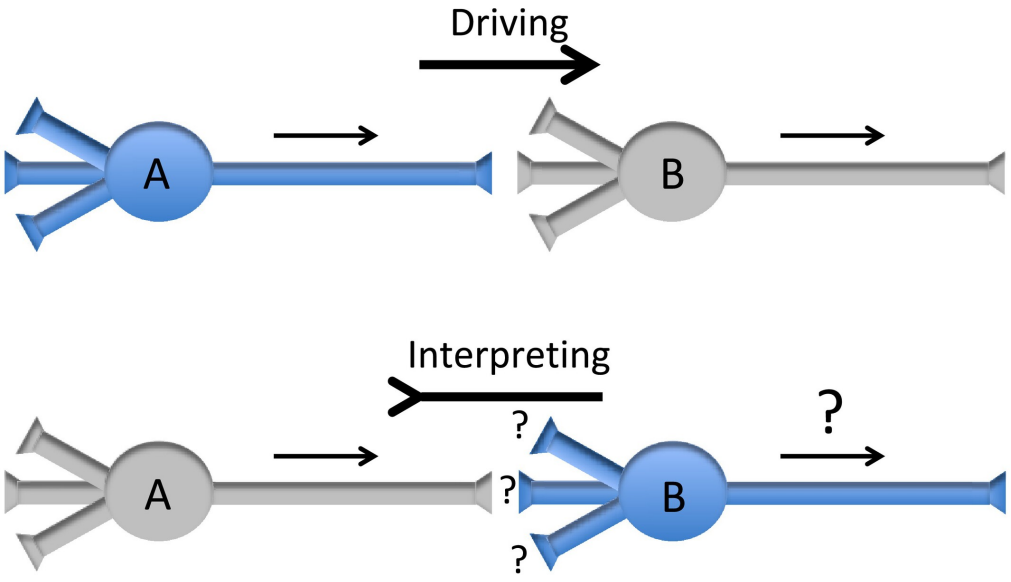

Figure 2 | Inverting the driving metaphor. The top row shows a driving relationship between neurons A and B, where B is effectively a passive element—activity in A *drives* activity in B. The bottom row inverts this relationship, highlighting the active role that neuron B plays in *interpreting* its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

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 behaviour 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, we will argue, 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 speaks to the need for a more expansive repertoire of causal concepts, harkening back to the sort of causal pluralism proposed by Aristotle.

3.5. 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, 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 force it imparts. 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) (see **Section 3.8**).

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 (1963) principles of ethology, 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 1600s 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 will see, 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 & 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* within that particular context. As Alex Gomez-Marin (2017) 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" (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*.

3.6. 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 such 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 new functionalities at higher levels (Juarrero, 1999, 2023; Hooker, 2013; Raja & Anderson, 2021; García-Valdecasas & Deacon, 2024; Ross et al., 2025). To see this, we have to go beyond a simple snapshot perspective on the system in order to recognise that the way in which the organisation of a system constrains the possibility space of what happens within is actually *enabling* certain phenomena to occur, certain tasks to be performed, and certain (emergent) global or macroscopic properties to obtain, that would otherwise be impossible. As Winning and Bechtel describe: "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 & Bechtel, 2018, p.307, our emphasis).

This idea of enabling constraints 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 in fact 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 & 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 (Prigogine & Stengers, 1984; Ismael, 2011; Juarrero, 2023). 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 feed back to 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., 2025). 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., 2025).

Another kind of self-organising dynamic is evident in the global states and trajectories of activity observed *within* 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 (Miller, 2016; Ebitz & Hayden, 2021; Durstewitz et al., 2023). 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 & Chemero, 2013; Ross et al., 2025).

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 (Gallego et al., 2017; Robson & Li, 2022; Durstewitz et al., 2023).

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 (Tinbergen, 1963; Marr & Poggio, 1976). 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 behaviour to occur?*' He refers to this as a structuring cause.

3.7. 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 (1988) 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 behavioural 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." (p.42, 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 (Seibt, 2016; Meincke, 2018; Nicholson & Dupré, 2018).

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 and 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 within 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 (2021) 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”. (pp.370–371).

3.8. 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 (Gallego et al., 2017; Semedo et al., 2019; Ebitz & Hayden, 2021). In the population-coding paradigm, it is these higher-order patterns that are thought to carry causal weight within the system (Semedo et al., 2020; Barack & Krakauer, 2021; Ebitz & Hayden, 2021; Mitchell, 2023a, 2023b). 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* (Ellis, 2012; Hoel et al., 2013; Flack, 2017; Comolatti & Hoel, 2022; 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 speak 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., 2025), 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 3.6** and **3.7**—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* (Woodward, 2005; Menzies & List, 2010; List & Menzies, 2017; Sinnott-Armstrong, 2019), 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).

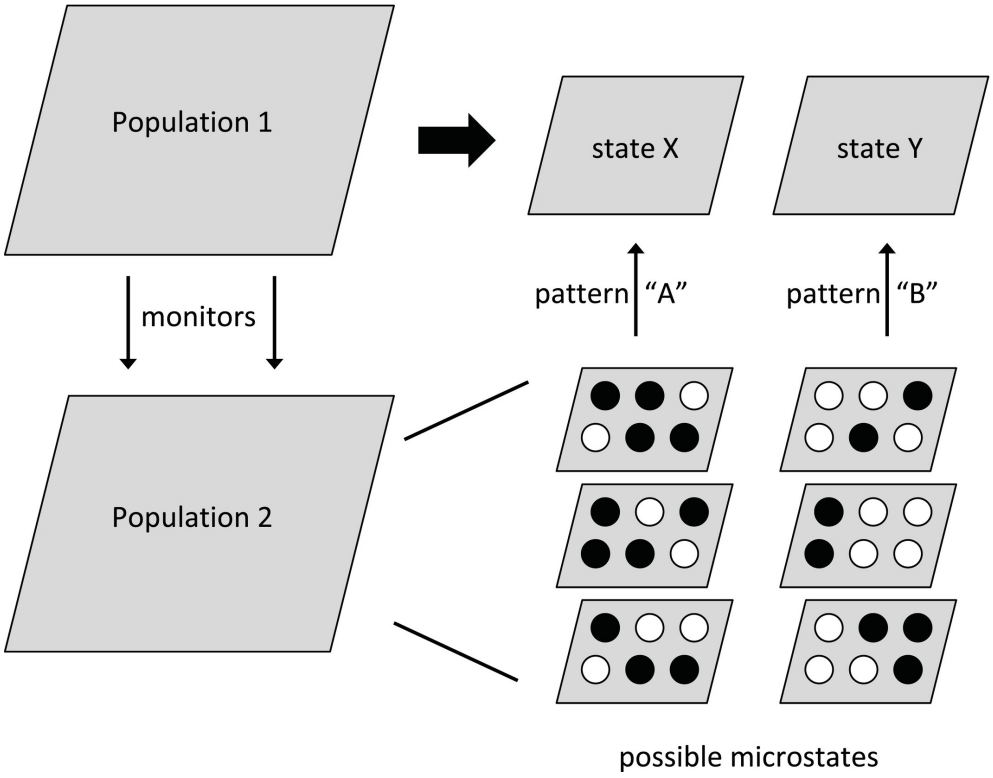

Figure 3 | Multiple realisability and macroscopic causation. A given population of neurons (Population 1) will monitor its inputs, depending on the criteria embodied in its afferent and internal connectivity. These criteria will determine the causal sensitivity and response to incoming macroscopic patterns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

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* 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*.

3.9. 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 (Friston, 2010; Clark, 2015). 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 & Lickliter, 2002; Pezzulo & Castelfranchi, 2007; Barsalou, 2008; Gopnik & Wellman, 2012). 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) (Piccinini, 2022; Mitchell, 2023b).

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 & 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. Of these patterns of neural activity, he wrote that they "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; Thiele & Bellgrove, 2018; Shine et al., 2021; Shine, 2023; Taylor et al., 2024).

Furthermore, 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'—in a way that depends on the meaning or content of mental states (discussed more below).

3.10. 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 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 affords control over the relevant 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.

3.11. 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.

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 *meaningful patterns of 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**).

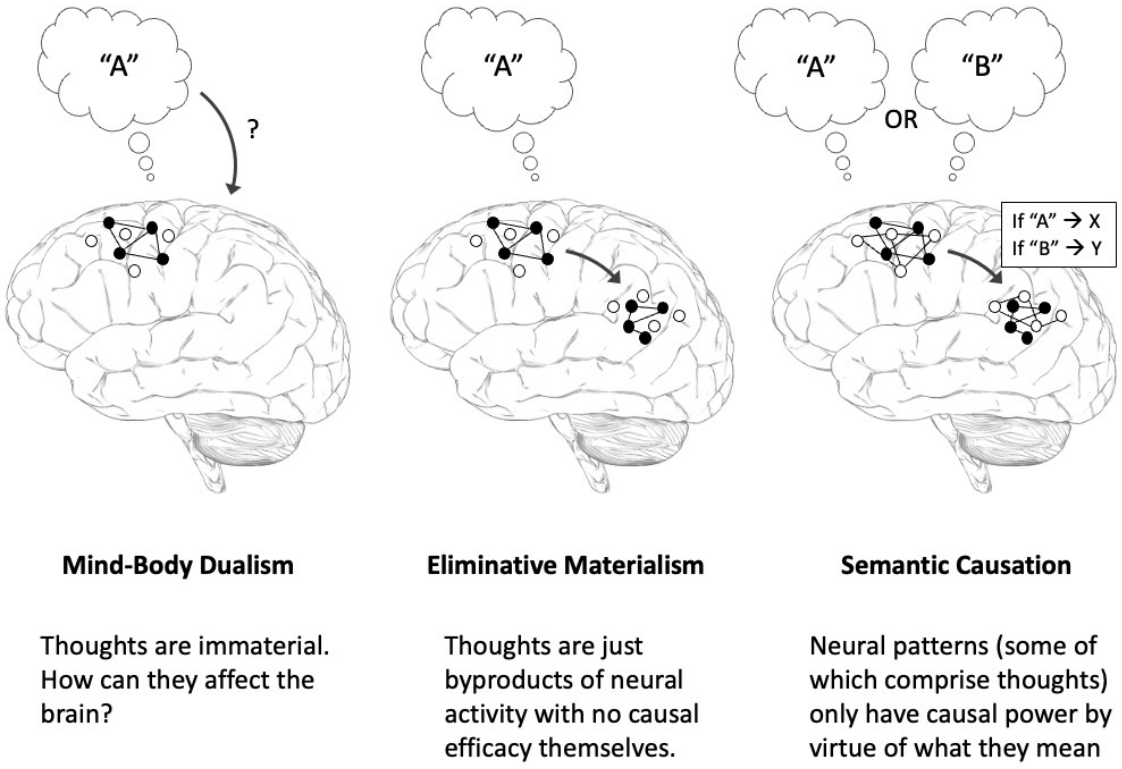

Figure 4 | Different conceptions of mental causation. Under a dualist viewpoint, thoughts are somehow non-physical, making it a mystery how they could affect neural processes. According to eliminative materialism, thoughts are *epiphenomenal*: all the causal work is done by their neural vehicles. Under a view of 'semantic causation' or "cognitive realism" (Mitchell, 2023a), thoughts are patterns of neural activity that only have causal efficacy within the system in virtue of what *mean* to the system *as a whole* (closely based on Mitchell, 2023a, with permission).

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.

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. (2022), say: **"**the events in the brain that 'cause' behavior"). Under the latter framing, the organism, as an agent, disappears from our 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 (**Chapter 2,** published as Potter & 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 behaviour'. 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 is best to do in any given scenario requires input from subsystems *across the whole brain* (Ismael, 2016; Mitchell, 2023a; see also **Chapter 5**).

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 & 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 & Mitchell, 2022; Mitchell, 2023a)—it is a causal agent unto itself. And our explanations ought to reflect that.

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 nervous 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.

Chapter 4

Reframing the Free Will Debate: The Universe is Not Deterministic

---

Status

Co-authored with George Ellis and Kevin Mitchell. This chapter is currently under review at the journal *Synthese*.

Author contributions

Manuscript primarily conceived, written, and edited by myself and Kevin Mitchell. Additional contributions to conceptualisation, writing, and edits provided by George Ellis.

"It ain't what you don't know that gets you into trouble.
It's what you know for sure, that just ain't so" (Mark Twain)

4.1. Abstract

Free will discourse is primarily centred around the thesis of determinism. Much of the literature takes determinism as its starting premise, assuming it true "for the sake of discussion", and then proceeds to present arguments for why, if determinism *is* true, free will would be either possible or impossible. This is reflected in the theoretical terrain of the debate, with the primary distinction currently being between compatibilists and incompatibilists and not, as one might expect, between free will realists and skeptics. The aim of this paper is twofold. First, we argue that there is no reason to accept such a framing. We show that, on the basis of modern physics, there is no good evidence that physical determinism (of any variety) provides an accurate description of our universe and lots of evidence against such a view. Moreover, we show that this analysis extends equally to the sort of 'indeterministic' worldview endorsed by many libertarian philosophers (and their skeptics)—a worldview which we refer to as determinism-plus-randomness. The paper's secondary aim is therefore to present an alternative conception of indeterminism, which is more in line with the empirical evidence from physics. It is this indeterministic worldview, we suggest, that ought to be the central focus of a reframed philosophy of free will.

4.2. Introduction

The thesis of determinism—"the thesis that there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.3)—plays a central and often defining role in the philosophy of free will. This is evident from the debate's current theoretical landscape. As Müller et al. (2019) write:

> "The free will debate standardly has its focus on the consequences of determinism for free agency. This is obvious when we look at the major distinction in the free will debate. One might expect that that distinction concerns theories that affirm freedom versus those that deny it. However, the major distinction among free will theorists is whether they are *compatibilists* or *incompatibilists* with regard to freedom and determinism. Thus, the central question that splits the debate into two opposing parties is about the implications of determinism for freedom" (p.6)

Or, as philosopher Peter van Inwagen (1983, p.55) puts it: "The main contested question in current discussions of free will is not, as one might expect, whether we *have* free will. It is whether free will is compatible with determinism."

In part, the reason why the theoretical terrain takes this form, and why free will discourse focuses so extensively on determinism, is historical. Some of the earliest known articulations in Western thought of the very idea that there might, in fact, be a *problem* of free will to contend with at all—that is, that there may be any reason to *doubt* the veridicality of our intuitive feeling of having free will—came within the context of various early forms of determinism. The early Greek materialist philosophers, Leucippus and Democritus, for example, formulated an influential atomistic worldview in which everything, including humans, was said to be ultimately made up of 'atoms' moving through 'the void' along necessitated pathways. The upshot, as Leucippus stated, was that 'nothing occurs at random, but everything for a reason and by necessity' (Edmunds, 1972). In other words, everything that happens was physically (and logically) determined to happen by what went before it. Later, Stoic philosophers added to this worldview the concept of laws of Nature (or laws of God), which they took to govern the trajectory of these atoms in the void, thereby providing an explanation for their deterministic nature (O'Connor & Franklin, 2022).

Philosophers of the time—including the Stoics themselves, as well as Aristotle and the Epicureans—recognised that such a deterministic worldview posed at least a *prima facie* threat to our intuitive free will. The Stoics, led most notably by the philosopher Chrysippus, sought to resolve the tension by arguing that what is needed for our actions to be 'up to us', in the sense required for free will, is just that their causes 'flow through us'—and hence our free will would not in fact be negated by the truth of determinism (a forerunner to the position now known as compatibilism). Aristotle and Epicurus, on the other hand, insisted that truly free actions could not be physically necessitated in this manner, and so argued that some form of indeterminism *must* exist in the universe if our actions are to be free (a forerunner to the position now known as incompatibilism) (O'Connor & Franklin, 2022). Whichever of these options one favours, the point for present purposes is that these early formulations of the problem of free will clearly took it to be *the problem of free will and determinism,* thereby laying the foundations for the conceptual terrain we see today.

In the intervening millennia, however, the discovery of quantum mechanics has led many philosophers, and most physicists (Schlosshauer et al., 2013), to officially reject the truth of determinism; few contemporary philosophers would now publicly subscribe to a thesis

of complete physical determinism. Yet it continues to play a central role in our philosophy of free will. Why is this? Why do we still dedicate the vast majority of our attention to whether free will is compatible with determinism, when most would outwardly reject determinism itself? There are perhaps a few reasons.

First, a small number of philosophers explicitly state that their focus on determinism stems primarily from a desire to safeguard their intuitions about free will, and—often more explicitly—moral responsibility, from certain potential future outcomes of scientific investigation. As John Martin Fischer and Mark Ravizza (1998) write:

> "Our contention is that … even if we discovered that causal determinism were true, there is a strong tendency to think that this sort of discovery should not make us abandon our view of ourselves as persons and morally responsible agents." (p.15)

On the contrary, Fischer (2007) later writes:

> "My basic views of myself and others as free and responsible are and *should be* resilient with respect to such a discovery about the arcane and "close" facts pertaining to the generalizations of physics." (p.45, *our emphasis*)

Free will philosophers of this persuasion might therefore continue to at least entertain the possibility of determinism for the purposes of their research, agreeing that such a thesis looks highly unlikely to be true of *our* world, but motivated by the belief that "[i]t is notoriously difficult to predict how future science will turn out… [so] it might be useful to have an answer to the question [of how we might retain our view of ourselves as free and responsible given determinism] in advance of the scientific issues getting sorted out" (Fischer et al., 2007, p.2).

Most philosophers take a different tack, however. They justify their focus on determinism on empirical grounds by appealing to some variant of the claim that, though strict *universal* determinism (i.e., a single, inevitable and pre-determined timeline of all events in the universe for all time) might not be true, something close enough *is* deemed by physicists to be true. Indeed, Fischer and Ravizza (1998) *also* make appeal to this line of reasoning themselves, stating that:

> "There is an additional reason to focus our attention on causal determinism. Although contemporary physicists tend to believe that causal determinism is false, they believe that something very much like it is true: a doctrine we shall call 'almost causal determinism.' On this view, macroscopic events are not, strictly speaking, causally determined, but they are very close to being determined" (p.15)

In this way, the thesis of determinism persists in the contemporary free will debate, rarely in its strongest form (which we will call 'universal determinism'), but more commonly in some weaker form (which we will call 'near-determinism'—although it has been variously referred to in the literature to as "almost causal determinism" (*ibid*), "neural-level determinism" (Pereboom, 2022, p.8), "macro determinism" (Honderich, 2001, p.465), "hard-enough determinism" (Caruso, 2012, p.4), and "for all practical purposes" determinism (Kane, 2001, p.8)).

There are two main forms of near-determinism. The first is the thesis of classical determinism. This is the view that indeterminacy does indeed exist at quantum scales of reality, but that it is of no relevance whatsoever to the problem of free will—either because the indeterminism of the quantum world completely disappears as subatomic particles combine to form the everyday objects of the macroscopic (or "classical") world, or because any quantum effects that *do* survive to the classical limit are so small and rare as to be entirely irrelevant to the processes of neural decision making and action selection. As philosopher Derk Pereboom (2022) puts it, classical determinism (or, for him, "neural-level determinism") is the view that "quantum micro-indeterminacies… are ordered with enough redundancy so that at the neural level, indeterminacy all but vanishes" (p.8).

In other words, the thesis of classical determinism is the view that, regardless of what is going on in quantum mechanics, the macroscopic world of brains, bodies and behaviour still evolves deterministically. It states that for objects and systems at this level of reality, which includes the putative subjects of free will, it remains the case "that there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.3).

There is also a second, weaker version of near-determinism at play in the free will literature, which we will call 'determinism-plus-randomness'. On this view, quantum indeterminacies are *not* seen as irrelevant to goings-on at the neural level. It is *not* taken to be the case that "indeterminacy all but vanishes" at macroscopic scales of reality. On the contrary, it is assumed that quantum effects may occasionally get amplified within the brain, introducing a non-negligible element of randomness into the macroscopic chains of causation involved in decision-making and action, and thereby rendering the subsequent decision and action outcomes causally *undetermined*.

Pereboom (2022) articulates this perspective nicely when he describes how:

> "[f]or alternative possibilities [in decision-making] to be significantly probable, there would have to be mechanisms that facilitate the "percolating up" of significant microlevel indeterminacies to the neural level, on the analogy of a

> Geiger counter that senses microlevel events and registers them at the level of the moving of a macrolevel indicator." (p.8)

This 'percolating up' model is perhaps the dominant way that indeterminism is conceptualised within the free will literature, common even among libertarian philosophers who seek to reject the thesis of determinism. And yet it is an implicitly near-deterministic worldview. By endorsing a picture in which quantum indeterminacies *occasionally* 'percolate up' to disrupt the deterministic processes unfolding at macroscopic scales, one is tacitly accepting that, for the most part (e.g., on local timescales), determinism holds true in the brain. That is, one is accepting determinism as "Nature's default mode" (Earman, 2008, p.817) at the level of neural activity, even in spite of the occasional possibility for disruption 'from below'.

Such a view is clearly evident in McKenna and Pereboom's (2016, p.16) description of how an "indeterministic" world (or, what we are proposing to call, a 'determinism-plus-randomness' world) is supposed to differ from a (universally or classically) deterministic one. They ask their reader to:

> "Consider the following model of a deterministic world, *Wd*, followed by a model of an indeterministic world, *Wi*. Let "e" represent an event. Let "—" represent a causally deterministic (d) relation between events. And let "…" represent an indeterministic (i) relation between events. Now consider each world*:*
>
> *(Wd): e1—e2—e3—e4—e5—e6—e7—e8, and so on with only d relations*
>
> *(Wi): e1—e2 … e3—e4—e5—e6—e7—e8, and so on with only d relations*"

On this model, indeterminism is conceptualised as something that gets intermittently *added* into an *otherwise deterministic* causal chain of events. It is for this reason that we refer to this popular worldview as 'determinism-plus-randomness', and treat it as a species of near-determinism, rather than viewing it as a *rejection* of determinism itself (as many of its proponents do).

These three metaphysical perspectives—universal determinism, classical determinism, and determinism-plus-randomness—collectively dominate the range of worldviews currently under consideration in the (naturalistic) free will literature. Here, we argue that none of them is supported by the empirical evidence.

In so doing, we hope to encourage a fundamental re-framing of the free will debate toward something that is more aligned with modern physics in two key respects: first, in its comprehensive rejection of physical determinism; and, second, in the way that indeterminism gets conceptualised within the debate. Importantly for free will philosophy,

we aim to show that the empirical evidence supports a picture of indeterminacy that does not merely present as random *additions* to an otherwise deterministic universe, but rather as a *pervasive indefiniteness* in the current states and in the future evolution of physical systems. What this means for the putative subjects of free will is that the future, as a whole, is simply *under-determined by its current state*; alternative possible futures, from the perspective of physics, are both inevitable and innumerable.

We end by proposing that, under such a worldview, the important question for free will philosophy is no longer 'where does the freedom come from?' but rather 'where does the control come from?' The appropriate focus is not on whether an agent 'could have done otherwise' than they did; but, rather, on how they managed to prevent all of the other physically possible 'otherwises' from happening, such that they were able to do what they wanted to do or what they *chose* to do. Reframing the free will debate in this way opens up a series of new problems, new questions, and a new way of conceptualising this age-old debate, which, we hope, can support future progress on this vexed issue.

In **Section 4.3**, we lay out the evidence and arguments against universal determinism, in particular by showing that the processes governing the evolution of quantum systems are genuinely indeterministic. In **Section 4.4**, we turn our attention to classical determinism. We consider a number of arguments claiming that, despite the existence of fundamental quantum indeterminacies, macroscopic (or 'classical') systems nevertheless evolve according to deterministic laws and processes. We show that none of these arguments is well supported and that the idea that classical physics is deterministic is not in fact a *result* of physics, but merely a convenient idealisation. Moreover, we present additional evidence that positively argues *against* classical determinism, and conclude that not only is there no good reason for philosophers to start with this premise, there are strong reasons not to.

In **Section 4.5**, we then consider the view of indeterminacy—"determinism-plus-randomness"—which is often taken as the starting point for libertarian arguments. This view takes deterministic evolution of physical systems as the default, but allows that it is occasionally interrupted by isolated random or undetermined events. This poses some well-known problems for libertarian views, as it is not easy to see how agents can be said to be in control of any action or choice that results from such an undetermined process, where what happens is ultimately just settled by the outcome of random microscopic events within them. Based on the evidence surveyed in **Section 4.4**, we present an alternative view of indeterminacy, which we refer to as "pervasive indefiniteness". Under this view, the future states of certain complex systems, such as ourselves, are *under-determined* by their current state and microscopic laws. In **Section 4.6**, we consider the

implications of this physical view for the free will debate, arguing that it calls into doubt the practical relevance of many of the debate's traditional positions and dividing lines. Lastly, in **Section 4.7**, we examine and respond to a number of possible objections to our arguments.

4.3. The Argument Against Universal Determinism

Universal determinism is the strongest version of the deterministic thesis. It states that "there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.3). Or, in slightly more specific terms, it is "the view that, for any given time *t*, a complete statement of the facts at *t*, together with a complete statement of the laws of nature, entails every truth as to what happens after *t*" (Palmer, 2014, p.4; also Fischer & Ravizza, 1998, p.14).[12]

This was the original version of the thesis, implied by the atomistic worldview of the early Greek materialists and the Stoic philosophers. It is also the version of the thesis articulated in the influential and evocative metaphor of Laplace's demon, an omniscient being for whom the state of the universe across all moments of time would be laid bare at once. Laplace describes the demon as:

> *An intelligence which, for one given instant, would know all the forces by which nature is animated and the respective situation of the entities which compose it, [and] if besides it were sufficiently vast to submit all these data to mathematical analysis, would encompass in the same formula the movements of the largest bodies in the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.* (Laplace, 1814, pp.3-4, translation from van Strien, 2014).

Some philosophers still accept the notion of universal determinism implied by the conceivability of such a demon, and many others are at least willing to entertain it as the premise for arguments about free will and moral responsibility. For example, though they differ on its implications, Gregg Caruso and Daniel Dennett, in a 2021 book, start out by both accepting the premise that: "facts about the remote past in conjunction with laws of nature entail that there is only one unique future" (p.5).

[12] Note that such determinism could only ever apply to a *closed* system. For the purposes of argument, we will assume (as proponents of determinism do) that the entire universe *can* be viewed as such a system. However, it should be noted that some models of spacetime, such as anti-de Sitter space (Hawking & Ellis, 2024), are not single closed systems.

We argue here that this premise is not sustainable; the universe is not deterministic in that fashion, either at quantum or classical levels. We start with the quantum level.

4.3.1. Quantum physics is not deterministic

The universal determinism worldview entails, at least in principle, "the claim that all processes can be fully described through a set of fundamental laws of nature [typically modeled using sets of differential equations], which always have a unique solution for given initial conditions" (van Strien, 2021, p.2). The thesis therefore relies on two key assumptions, both of which were undermined in the twentieth century by the discovery of fundamental indeterminacy in quantum mechanics.

Universal determinism's first critical assumption is that systems (including whole universes) exist in precisely defined (or hypothetically *definable*) states at every given instant of time. If they did not, then we would not be able to specify the initial conditions of the system from which it uniquely (i.e., deterministically) evolves. That is, there would be no physical reality that corresponds to the claim that Laplace's Demon could "know all the forces by which nature is animated and the respective situation of the entities which compose it".

However, this assumption is straightforwardly ruled out by the Heisenberg Uncertainty Principle (HUP). The HUP describes the fact that in a system of elements with wave-like properties (as is the case in quantum systems) various "conjugate variables", most famously including the position and momentum of particles, are related to each other by a Fourier transformation. What this means is that the more precisely one variable is defined, the less definition the other variable has. In other words, the HUP tells us that it is physically impossible to give 'a complete statement of the facts at *t*' for the entire universe—as is a stated requirement of universal determinism (Fischer & Ravizza, 1998)—because 'completeness' in one area of the system *necessarily* comes at the cost of 'completeness' in another.

It should be noted that the use of the word "uncertainty" to describe this principle can be misleading as it may give the impression of a purely epistemic phenomenon. That is, it may give the impression that the Uncertainty Principle refers to the inability of an observer to precisely *measure* both variables at once. However, the principle (originally dubbed the Indeterminacy Principle by Heisenberg) is emphatically ontic in nature. Conjugate variables simply cannot simultaneously exist with infinite precision, meaning it is physically impossible for a system—or the universe as a whole—to exist in an infinitely precise and exhaustively definable state at any given time.

This is demonstrated in a concrete way by the observation of "zero point energy". When the temperature of a system is reduced to absolute zero (the absolute lowest it can go), one might reasonably assume, given that the motion of particles is what generates temperature in the first place, that the motion of the system's constituent particles would come to a standstill and that all energy in the system would be lost. This is not the case, however. "Zero point energy" is the observation that there always remains some energy in the system, even at absolute zero (Milonni, 1994). The explanation for this is that, if it were not the case, then the conjoint precision of the momentum and the position of particles (as possible manifestations of the pervading quantum fields) would be infinite, and that would violate the HUP. Instead, then, energy remains in the system due to the irreducible probability of quantum fluctuations occurring as a result of the HUP. The empirical evidence for zero point energy is thus consistent with the interpretation that the HUP is an ontic—and not merely epistemic—feature of our world.

The second key assumption of universal determinism is that any isolated physical system (and, indeed, the entire universe) evolves *necessarily* from one time-point to the next. Mathematically speaking, it is the assumption that the transitions between 'states' of the system are derived from (or described by) sets of differential equations which always have a unique solution for the given antecedent state. This is the notion of a deterministic law of nature, which is what ultimately underlies the universal determinist's claim that *if* Laplace's Demon had access to the position and momenta of every particle in the universe, it could then "submit all these data to mathematical analysis" and deduce "the [future] movements of the largest bodies in the universe and those of the lightest atom". The relevant assumption here being that such a 'mathematical analysis' would necessarily yield a *unique* solution (and, also, that such a mathematical formalism fully captures how the universe itself reaches the next state).

Again, this assumption is challenged by research in quantum physics. The evolution of quantum systems from one time-point to the next—i.e., the transition between 'states'—is fundamentally indeterminate, not only as a feature of the initial indefiniteness described by the HUP, but more generally, in that the state of a system is described by a *probability density function*, rather than with fixed, precise values for every parameter. The Schrödinger equation is what provides us with the means to calculate how such a quantum system will evolve through time, and it is often described as being deterministic. This is true insofar as *the equation itself* does not admit any randomness and theoretically has a unique (unitary) solution for any future time-point. However, the solution it gives is still a distribution of probabilities, not a single actuality. What the equation tells us is how the

probabilities of observing the system in one state or another will evolve over time, if the system is not observed. When we want to see how the physical system *actually* evolves, what we get in a single trial is a "collapse" to one of those possible states—a collapse that seems to be genuinely un(der-)determined by any antecedent conditions. (See **Section 4.7** on Possible Objections for comments on unitarity and the conservation of information).

There is no reason to think—as is sometimes suggested—that the resolution of this indefiniteness in the system relies on *an observer*, either. It arises any time some physical interactions force the system into a definite state. Empirically, across many such trials, the Schrödinger equation very accurately predicts the frequency of different outcomes (for simple systems at least). But in any given trial, the outcome seems to be genuinely probabilistic—a random draw from the probability distribution. In that sense, how the potentialities of a quantum system become the actualities that obtain and that we observe, in practice, does seem to be truly undetermined.

A variety of theories have been proposed to try to rescue determinism within such systems. These include ones that invoke hidden variables of one kind or another, which are taken as actually explaining (in virtue of causally determining) which specific outcome arises in any specific instance (reviewed in Earman, 2004). However, no evidence for such hidden variables has ever been found, and their existence seems to be ruled out by Bell's theorem (under locality) (Bell, 1964), which has been empirically supported time and again (e.g., Acín & Masanes, 2016; Abellán et al., 2018).

An alternative approach —dubbed the Many Worlds Hypothesis—is to simply assume that every possible value in the probability density function actually gets realised, in a process that generates separate universes for each possibility at each such event (Everett, 1957). While such a scheme can be made mathematically consistent, there seems little reason to take it seriously as a physical reality or even as a metaphysical possibility. In the first instance, it assumes that the entire universe can be described by a single Schrödinger equation, but there are strong arguments against this notion (Drossel, 2017; Ellis, 2023). However, more generally, the Many Worlds Hypothesis simply does no work in explaining the phenomenon in question—namely that, in the universe we are experiencing, only one of the possibilities actually occurs.

We take it that it is therefore generally accepted that there exists fundamental indeterminacy in quantum systems and that the evolution of quantum systems is

essentially non-deterministic.[13] The upshot, as many philosophers of free will seem to agree, is that the thesis of universal determinism is highly implausible as a description of *our* universe.[14]

4.3.2. New possibilities emerge as the universe expands

Before moving on to the more contested thesis of classical determinism, it is worth briefly pausing to mention the status of a claim that is often seen in the free will literature and that seems to go hand-in-hand with universal determinism. This is the claim that the conditions of the early universe, in the moments after the Big Bang, causally fixed everything that has happened and will happen in our universe. Or, in an alternative framing, it is the claim that, in some sense, the Last Scattering Surface of the early universe could have contained all of the information necessary for a hypothetical, Laplacean being to predict every future event and state—including the Cretaceous extinction event, the invention of iPhones, and every single choice and action an individual human will take in their lifetime. Here is philosopher Christian List describing (though not endorsing) such a claim:

> "A second kind of argument derives the unreality of free will from the claim that the laws of physics may be deterministic, meaning that the initial state of the universe, say at the Big Bang, pre-determined all subsequent events; so, there would be no room for alternative possibilities to choose from. You may think that you had a choice whether to read this article or not, but in reality, your decision was made for you by the world's initial conditions." (Caruso et al., 2020, p.2)

Thankfully this unsettling proposal has no actual basis in physical reality. First, the indefiniteness of quantum systems described above renders it both impossible *and* unnecessary—as explained by Georges Lemaître (1931), who first proposed the idea of the universe beginning from a singular quantum state:

---

[13] This would also mean that these systems are not time-symmetric either, meaning that information *really is* created and destroyed as time proceeds. We address this point further in the Possible Objections section below.

[14] Of course, how exactly to interpret the fundamental indeterminacy of quantum systems is far from settled. Not only are there attempts to interpret it deterministically (which we have suggested are unsuccessful), but different indeterministic interpretations also exist. Recent work by Jacob Barandes (2023; 2025), for example, demonstrates how one can fundamentally reformulate quantum theory "in the language of trajectories unfolding stochastically in configuration spaces", in a way that does not rely on superpositions or wave function collapses, and still recover the full range of quantum phenomena.

> "Clearly the initial quantum could not conceal in itself the whole course of evolution; but, according to the principle of indeterminacy, that is not necessary. Our world is now understood to be a world where something really happens; the whole story of the world need not have been written down in the first quantum like a song on the disc of a phonograph. The whole matter of the world must have been present at the beginning, but the story it has to tell may be written step by step." (p.706).

Second, Lemaître's view can be supported by an informational perspective. Because information requires a physical substrate (Landauer, 1991), the amount of information that can be held in any finite region of space is limited. In particular, it is limited by the Bekenstein bound, which caps the peak entropy of a system of finite size, and thus its capacity to hold information (Shannon & Weaver, 1949; Bekenstein, 2004). For a hypothetical Laplacean demon to be able to predict the invention of iPhones from the Last Scattering Surface of the early universe, however, this finite region of space would need to contain a near-infinite amount of information. Thus violating the Bekenstein bound and forcing us to conclude that such a proposal is indeed physically impossible.

Lastly, claims about the early universe pre-determining everything that we do are *also* undermined by the fact that the universe has continuously expanded since the Big Bang, creating an ever-expanding possibility space of states that it could be in (**Figure 5**). This rate of expansion outstrips the rate of equilibration, meaning that while entropy continues to increase, so too does information (Layzer, 1975; 2021). As Arthur Eddington said, in 1935: "The expansion of the universe creates new possibilities of distribution faster than the atoms can work through them." [15] (p.66) It is therefore not plausible that the outcomes of all future physical events could be pre-determined, *fixed* by, or encoded *in* the state of the universe at one point in time when the possibility space in which these futures will evolve does not yet even exist. As we will argue later on, this point applies not only to the origins of the universe, but also to the state of the universe at any subsequent time-point (including the present).

[15] On this view, the fact that the early universe was in a state of low entropy was simply because it was very small. This meets the requirements of the so-called Past Hypothesis (which postulates that the universe must have started in a notably low-entropy state) without any special pleading (Layzer, 2021).

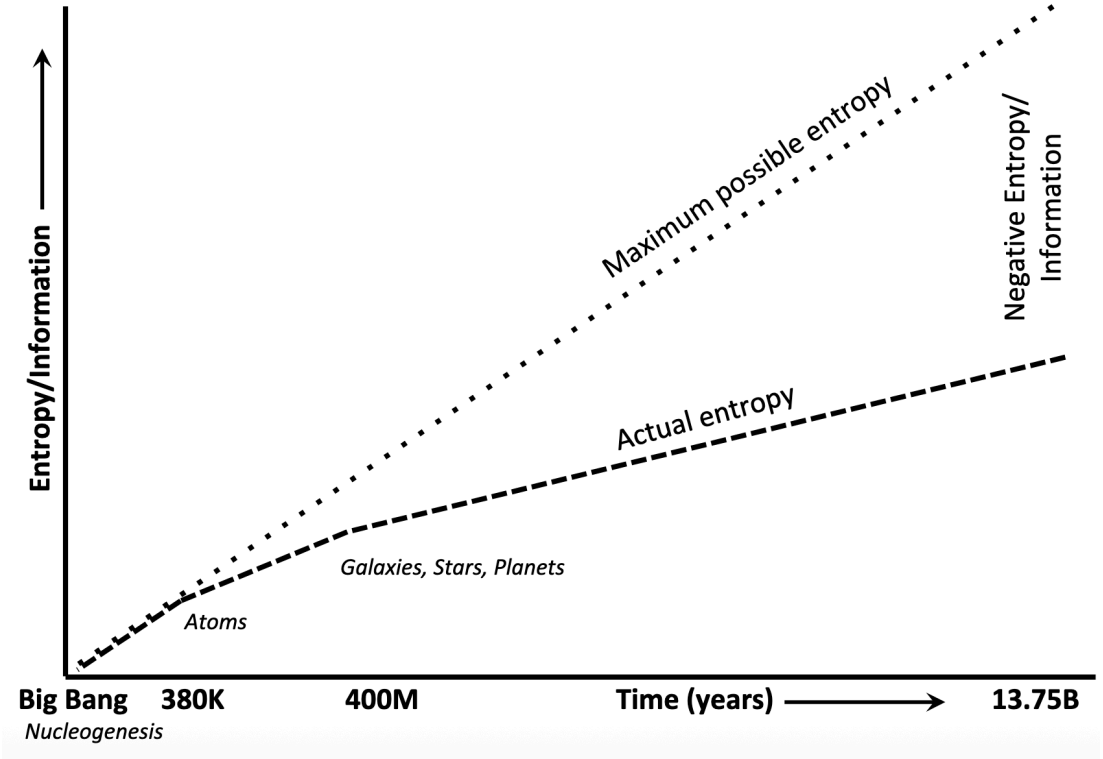

Figure 5 | The continuously expanding universe. At the Big Bang, the universe was in a state of very low entropy, relative to later times. As it expanded, and atoms, then galaxies, stars and planets emerged, the processes of physical equilibration lagged behind the pace of expansion of physical space (and possibility space). This means that even as entropy increased, so too did information. (Based on an illustration by Robert Doyle, depicting the scenario described by David Layzer in his 1975 article, *The Arrow of Time*).

Most philosophers of free will accept that the extremely strong evidence for fundamental indeterminacy at quantum levels ought to put the nail in the coffin of universal determinism. What this means for systems at the so-called "classical" level, where everyday macroscopic objects reside, is more contested, however. It is to this we now turn.

4.4. The Case Against Classical Determinism

Intuitively, it might seem obvious that, if determinism does not hold at the quantum level, then it cannot hold at the classical level, either. At least not in the strict sense in which, for macroscopic objects and systems, "there is at any instant exactly one physically possible future", if that future is taken to be defined with microscopic precision (van Inwagen, 1983, p.3). After all, everything is made up of atoms and subatomic particles and, ultimately, quantum fields. Quantum indeterminacy should thus be constantly at play in the deepest reality of every system, including those systems that may be the putative subjects of free will.

Yet, the relevance of quantum indeterminacy for goings-on in the macroscopic world is often dismissed by philosophers of free will (and by many physicists, too). These arguments lean on a number of different ideas, but they all lead to the conclusion that classical levels are somehow *insulated* from the indeterminate goings-on at quantum levels—a metaphysical view known as 'classical determinism'. If classical determinism is true, then:

> "even though indeterminism reigns in our brains at the subatomic quantum mechanical level, our macroscopic decisions and acts are all themselves determined" (Dennett, 2015, p.148)
>
> "although there is indeterminism at the micro-level, the level of small particles, there is still determinism at the macro-level, which includes neural events and everything with which we are ordinarily familiar" (Honderich, 2002, p.140)

In support of this conclusion, philosophers have provided various forms of what we will refer to as 'Las Vegas arguments' (since their objective is to ensure that what happens in the quantum realm stays in the quantum realm). Clearly, a metaphysics of classical determinism cannot just be assumed or asserted for the purposes of debate. A successful Las Vegas argument is required to motivate the assumption of classical determinism and its relevance for discussions of free will.

In the first part of this section, we argue that none of the proposed Las Vegas arguments are successful in their task. In doing so, our aim is to demonstrate that classical determinism is not something that has been conclusively proven and, as such, it is not a metaphysical constraint that naturalistic philosophers of free will *need* to abide by.

In the second part of the section, we argue further that, not only has classical determinism not been *proven* by physics, but in fact there are some very strong conceptual and empirical reasons to think that it is highly implausible as a description of the way that our macroscopic universe evolves. That is, it is not only the case that naturalistic philosophers of free will *need not* adhere to classical determinism in their work, but there are compelling reasons to think they *should not*.

4.4.1. The failure of Las Vegas arguments

Here we survey three flavours of Las Vegas argument and show why they do not hold.

a) Classical physics *just is* deterministic.

The simplest kind of Las Vegas argument simply rests on the (mistaken) belief that classical determinism has been empirically proven by physics. In this view, regardless of

how exactly it happens, the assumption is that we just *know* that the classical world obeys its own rules, untroubled by whatever random events may be occurring at quantum levels. We just *know* that "determinism should be thought of as Nature's default mode" when it comes to macroscopic systems such as ourselves (Earman, 2008, p.817).

How do we know this? Because determinism has apparently been shown by, or perhaps is a well-established *result* of, the best physical theories we have for describing the macroscopic world—namely, classical physics (which includes Newtonian mechanics and Maxwell's electrodynamics). Indeed, the view that the classical world is deterministic has held sway among physicists since Newton's development of his laws of motion, which proved so effective in predicting the orbits of the planets. It might therefore seem reasonable for free will philosophers to extrapolate from Newton's success in this domain to the idea that all systems must be similarly deterministic, at least at macroscopic levels of resolution, even if their complexity makes them unpredictable in practice. As the common refrain goes, 'the (classical) universe is clearly deterministic enough for us to have gotten to the moon'.

But is it actually right to say that classical physics has *proven* that the macroscopic world of brains, bodies and behaviours evolves deterministically? Or have we perhaps been misled by an exceptionally obliging solar system, as was suggested by the philosopher Elisabeth Anscombe (1971, p.99) who said that: "[t]he high success of Newton's astronomy was in one way an intellectual disaster: it produced an illusion from which we tend still to suffer"? In other words, might it in fact be the case that the widespread belief that classical systems are provably deterministic is one of those ideas, to use Mark Twain's phrasing, that we "know for sure, that just ain't so"?

Philosopher and historian Marij van Strien (2021) expertly addresses these questions through a historical lens, asking: "*was physics ever deterministic*?". After surveying the extent to which physics in the pre-quantum (now referred to as 'classical') period was committed to the thesis of determinism, she concludes:

> "during the period which we now describe as classical, determinism was not so much an established result of physics, but rather an expectation, and … during the late nineteenth and early twentieth century, it more and more took the form of a methodological principle or necessary presupposition of science, rather than an ontological claim." (p.2)

Most importantly for our purposes, van Strien explains that the thesis of determinism, as it was originally articulated within the modern context (e.g., by Laplace), was deeply

embedded within a very specific research program within classical physics—one that sought to reduce all of physics to mechanics (motion), and all of mechanics to the motion of idealised point-mass particles. Yet, even before the discovery of quantum physics at the turn of the twentieth century, these explanatory objectives had mostly been abandoned—due, among other things, to the implausibility of a solely point-mass conception of matter and a growing disbelief that statistical laws (such as the second law of thermodynamics) could ever successfully be reduced to (or derived from) the laws of mechanics (see also Drossel, 2015). Consequently, as van Strien says, "by the end of the nineteenth century, it had become increasingly unfeasible to reduce all of physics to a basic set of equations and to establish that these equations would always have a unique solution for given initial conditions" (p.10). In other words, the original basis for the claim that (classical) physics entails a deterministic picture of the macroscopic universe had become fundamentally implausible.

Instead, the author notes, determinism took on the role of a methodological presupposition within classical physics—an explicitly pragmatic assumption that was seen as necessary for scientists to *do* science. Many physicists in this period, including Max Planck, Ernst Mach and Henri Poincaré, publicly adopted a position of agnosticism with regard to whether the (classical) universe was *really* deterministic or not, noting only that determinism was a necessary heuristic principle for guiding their work. As Poincaré put it, "we are determinists voluntarily" (1921, p.347).

Of course, the fact that determinism is not, and has never been, an established *result* of classical physics does not, by itself, imply that the macroscopic world is *not* deterministic. As many have noted, proving determinism would be effectively impossible. It does, however, undercut the impression (common in philosophy) that one can safely *assume* that neurons, brains and bodies are causally insulated from quantum effects. Instead, this is something that needs to be argued for or defended. We will now turn our attention to some such attempts to defend classical determinism.

b) Decoherence quarantines quantum indeterminacies.

A more mechanistic Las Vegas argument appeals to notions of quantum decoherence in the so-called quantum-to-classical transition (Zurek, 1991; Schlosshauer, 2005, 2014). Decoherence is an established phenomenon in physics which describes the elimination of quantum interference effects when a quantum system interacts with its environment. The consequence is that the system starts to behave in a more classical fashion. For free will philosophers, this has sometimes been taken to imply that "quantum effects could just… be self-canceling" (Dennett, 2015, p.148) or that "micro-indeterminacies 'cancel each other

out,' and we get macro-level determinism" (McKenna & Pereboom, 2016, p.23). Decoherence therefore seems to provide a potential argument for classical determinism that goes beyond a mere (and mistaken) appeal to an empirical *proof* of classical determinism and provides, instead, a mechanism *whereby* classical levels could become insulated from quantum indeterminacy.

Such an argument is not successful, however. The claim that decoherence resolves the quantum "measurement problem" (the fundamentally *random* emergence of actualities from a probability distribution) is not, in fact, supported (Adler, 2003; Schlosshauer, 2005, 2014; Drossel & Ellis, 2018).

> "If we understand the 'quantum measurement problem' as the question of how to reconcile the linear, deterministic evolution described by the Schrödinger equation with the occurrence of random measurement outcomes, then decoherence has not solved this problem." (Schlosshauer, 2014, p.19)

On the contrary, any superpositions—i.e., the range of possible states that the system *could* be observed in—that existed *or that arise* in its microstates must still be resolved, apparently still at random. Decoherence simply means that there should not be superpositions of macrostates (like simultaneously live and dead cats, as in Schrödinger's famous thought experiment). Moreover, each resolution of quantum uncertainty will generate a new state whose future evolution is described by *a new (probabilistic) wave function*, which will in turn resolve (randomly) into some actuality, with another new wave function, and so on. If we take seriously the notion that quantum fields and systems really are the lowest level of physical reality, not just under special conditions of isolated streams of electrons or photons in the laboratory, but in more complex systems, then we must accept the constant introduction of some level of noise or randomness, not just in a once-off quantum-to-classical transition which permanently eliminates all quantum nature from the macroscopic system, but as an ongoing process.

As Sean Carroll (2024, p.36) describes:

> "The rule is this: whenever we measure an observable, whatever the wave function was before the measurement, it immediately collapses onto some definite value of the quantity being observed. The new post-collapse wave function then evolves according to the Schrödinger equation, until it is observed and collapses again."

Thus, the particular mechanism of decoherence (which itself remains poorly understood and open to interpretation) simply does not support a Las Vegas-style argument for classical determinism. Quantum events do not 'cancel each other out' in a way that

eliminates all presence of randomness in the deepest recesses of our being. Indeterminism is instead an ever-present feature of the fundamental matter out of which we are made.

c) Classical determinism is statistical and emergent.

A third kind of Las Vegas argument *accepts* that there is randomness constantly at play at quantum (or just microscopic) levels within the system, but claims that none of this actually matters for how the macroscopic variables and properties of the system will evolve over time, because all of this low-level indeterminacy simply gets averaged out or coarse-grained over, leading to deterministic dynamics at the macroscopic level. It is therefore essentially an argument akin to the mathematical law of large numbers, which states that the average of a large number of random events typically converges onto a stable, regular or 'true' value. As Philip Ball (2006) puts it, it "is a way of saying that pure randomness gives way to determinism if the number of random events is large" (p.76). So, just as a large collection of simultaneous coin flip events reliably produces a stable and predictable outcome at the statistical level (namely, a Bernoulli distribution), so too—this version of the Las Vegas argument says—does all of the microlevel randomness within a system produce a fixed and determinate 'average' overall effect at the macrolevel.

The result is that the macroscopic system can be viewed as evolving strictly according to the (deterministic) laws of classical physics, *when observed at this macroscopic scale of resolution*. Philosopher Daniel Dennett (2015) gestures at this type of Las Vegas argument when he asks us to consider a robot living in what he describes as a 'deterministic world':

> "Once again we will take the world of the robot explorer, for then we can know just what we are stipulating in saying that its control system is completely deterministic: we design it to be deterministic, to be highly resistant to micro-level noise and random perturbation. (It has no built-in Geiger counters to propagate random effects; it is designed instead to damp out such effects.)" (p.126)

For Dennett, the robot evolves deterministically *despite* the continual presence of microscopic randomness, because it is able to 'damp out such effects'. Indeed, this is how most engineering systems work. Another example might be the fact that the temperature of a gas at equilibrium (a macroscopic variable) does not change *despite* the microscopic fluctuations of its individual molecules.

The result is a kind of statistical or *emergent determinism* at the classical level, where the macroscopic dynamics are taken to be insensitive to the detailed fluctuations happening at the lowest levels and can still evolve in predictable, deterministic ways *at a certain level of resolution*. Note the difference from the idea of universal determinism, which is a bottom-

up argument, the idea being that everything that happens at macroscopic levels is deterministic *because* all the low-level processes are deterministic. Here, the idea is that the highest levels can have some deterministic dynamics unto themselves, *despite* all kinds of possible noise at the lowest levels.

Is this argument successful in proving that classical determinism holds in our universe? It certainly seems to accurately describe something like the orbits of the planets, which are incredibly well described by the laws of classical mechanics, despite the fact that, say, the atmospheric conditions on any of those planets evolve non-deterministically. Those details just do not affect the planet's orbit (at least not to any extent that we would ever possibly need to care about). But can we assume that behaviour observed for such simple systems can be extrapolated to more complex ones? Would free will philosophers be justified in claiming that, because planetary solar systems exhibit an emergent determinism, we can assume that brains do too? We argue in the following sections that this inference is not valid; in many classical systems (even the solar system itself), macroscopic dynamics are highly sensitive to the ways in which microscopic indeterminacies are resolved.

To sum up this section, these varieties of Las Vegas arguments fail to make the case for the insulation of the classical realm from quantum goings-on, *in toto*. Classical determinism is not, and never has been, an established *result* of physics. It is not a worldview that has been proven by classical physics or one that has been secured by the discovery of quantum decoherence. There is perhaps some evidence that a sort of emergent determinism *can* take hold in simpler classical systems; however it does not follow from this that it does so in more complex systems like the human brain and body. Hence, in conclusion, there seems no empirical requirement for philosophers of free will to accept the premise that "although there is indeterminism at the micro-level, the level of small particles, there is still determinism at the macro-level, which includes neural events" (Honderich, 2002, p.140). In the next section, we go one step further by arguing that, for a certain class of macroscopic system, which includes the human brain, there are also some strong, positive reasons to think that their evolution is decidedly *not* deterministic in the manner prescribed by classical determinism.

4.4.2. Direct arguments against classical determinism

Clearly, the classical determinism worldview depends on the same set of assumptions as the universal determinism worldview, modified slightly for its narrower domain of applicability. To recap, these assumptions are: (i) that (macroscopic) systems exist in precisely defined (or hypothetically *definable*) states at every given instant of time, and (ii) that the transitions between these 'states' of the (macroscopic) system are derived from

(or described by) sets of differential equations which always have a unique solution for the given antecedent state (i.e., that the laws of classical physics are intrinsically and unfailingly deterministic).

The first of these assumptions could presumably be cashed out in one of two ways. First, when one speaks of the 'state' of a macroscopic system, such as a brain, one could be referring to the collective states (position, momentum, etc.) of all of the individual atomic or subatomic elements that materially constitute the system at that time. On such a view, for the 'state' of a macroscopic system to be precisely defined just *is* for each of its smallest (i.e., quantum) constituents to exist in a precisely defined state. However, as we saw in **Section 4.3.1**, such a situation is physically ruled out by the HUP.

An alternative interpretation of assumption (i) is the more commonplace view that each of the classical-level physical parameters of a macroscopic system (e.g., its position, length or centre of mass) exists in a precisely defined state, at any given time. Typically, this would mean that each of these parameters could, at least in principle, be described with infinite precision using a real number—that is, where all the decimal places are given, all at once (although cf. Kwok, 2020). We take it that this is generally what is required for a macroscopic system to exist in a precisely defined (or hypothetically definable) state. Yet, this is a situation that cannot actually hold in the physical world. Such precision is simply a mathematical idealisation. As expressed by pioneering quantum physicist Max Born (1969): "*Statements like 'a quantity x has a completely definite value' (expressed by a real number and represented by a point in the mathematical continuum) seem to me to have no physical meaning*" (p.81). Likewise, Karl Popper (1950) said: "*infinitely precise and complete knowledge is also 'in principle' unattainable*" (p.123).

To see why, consider again the argument presented in **Section 4.3.2** for why the initial universe could not contain sufficient (i.e., infinite) information to predict every future state and event. According to this argument, such a proposition would require a finite region of space to hold an infinite amount of information—thereby violating physical law. As compellingly argued by physicists Flavio del Santo and Nicolas Gisin (Del Santo & Gisin, 2019; Del Santo, 2021; Gisin, 2021a, 2021b), the same logic applies to any subsequent time-slice of the universe, and even more so to particular systems *within* it. Such systems occupy finite space. Yet, for the physical parameters of these systems (at both micro- and macro-scales) to exist in the precisely defined state demanded by classical determinism, it would need to be the case that these parameters can be described (at least in principle) using real numbers. Real numbers, however, typically feature no structure at all (i.e., they are fundamentally incompressible) and thus contain an infinite amount of information. It

would therefore be a simple violation of physical law for the parameters of a macroscopic system to *actually* exist in a precisely defined state at any given instant of time.

These authors show, therefore, that the supposed determinism of classical physics ultimately rests on what is really just a mathematical idealisation, with no plausible basis in physical reality; namely, the idea that the real numbers which describe the physical parameters of any system are given with infinite precision, all at once. Indeed, even David Hilbert, the most ardent defender of the mathematical concept of infinity, admitted that this concept had no basis in physical reality (Ellis et al., 2018): "The infinite is nowhere to be found in reality, no matter what experiences, observations, and knowledge are appealed to." Or, as Nicholas Gisin (2021b) puts it, '*real numbers are not really real'.*

The upshot is that there seems to be no way, in principle, for the state of a classical system to ever be defined (or, indeed, *exist*) with sufficient precision to meet the job description of determinism. The so-called 'initial conditions' of the system are always going to be somewhat fuzzy, vague and undefined; they exist in well-defined states up to a certain degree of resolution, but at higher levels of precision their state is truly indeterminate (Mariani & Torrengo, 2021). That is, after a certain number of decimal places, the system's state will just be undefined or indefinite (Ben-Yami, 2020).

Under this view, classical determinism becomes fundamentally impossible as an ontological claim—an artefact of our mathematical idealisations, rather than a feature of our world. As physicist Barbara Drossel (2015) explains:

> "In classical mechanics, the state of a system can be represented by a point in phase space. The phase space of a system of N particles has 6N dimensions, which represent the positions and momenta of all particles. Starting from an initial state, Newton's laws, in the form of Hamilton's equations, prescribe the future evolution of the system. If the state of the system is represented by a point in phase space, its time evolution is represented by a trajectory in phase space. However, this idea of a deterministic time evolution represented by a trajectory in phase space can only be upheld within the framework of classical mechanics if a point in phase space has infinite precision. If the state of a system had only a finite precision, its future time evolution would no more be fixed by the initial state, combined with Hamilton's equations. Instead, many different future time evolutions would be compatible with the initial state." (p.2)

Moreover, if "many different future time evolutions would be compatible with the [same] initial state" then the second key assumption of classical determinism also seems under

threat. This assumption required that the time evolution of a macroscopic system proceeds *necessarily* from one time-point to the next, governed by some sort of deterministic causal laws. However, as Drossel explains, such a worldview would be highly infeasible if, as we have argued above, the initial conditions of the system cannot be specified with infinite precision. For such systems, "the state [of the system always] has a fundamental indeterminacy that leads to an indeterministic dynamics" (Del Santo, 2021, p.67). This is most noticeably true for systems with "chaotic" dynamics, such as the brain, as we explore in the next section.

4.4.3. Chaotic systems

The importance of this fundamental imprecision or indefiniteness for the problem of free will becomes especially acute when we consider that biological systems are inherently chaotic, dominated by non-linearity and a sensitivity to initial conditions (Deco & Kringelbach, 2020; Terada & Toyoizumi, 2024; Deco et al., 2025). This sort of sensitivity was famously discovered by Lorenz in running computer simulations of a simple weather system (Lorenz, 1963; 1969). He found that re-running a simulation from a given point, but with the initial parameters truncated after fewer decimal points than in the first run, produced a trajectory of the system that initially followed that of the first run closely, but that then began to diverge, such that, after some period of simulated time, there was no longer any correspondence at all between the two trajectories. That is, the evolution of the system as a whole was exquisitely sensitive to tiny changes in the starting parameters, to the point that, beyond a certain time horizon, two systems that differ only fractionally in their initial conditions "will evolve into two states differing as greatly as randomly chosen states of the system" (Lorenz 1969, p.289; see also Palmer et al., 2014). Such systems became known as chaotic systems.

It is often claimed that, despite being unpredictable in practice, the evolution of these chaotic systems is still nevertheless deterministic. That is, their future *is* actually fixed by their initial conditions; it is just that *we* cannot know what those are in full detail (i.e., the observed indeterminacy is merely epistemic in origin). This kind of deterministic situation may in fact hold for computer *simulations* of such systems, where any given set of initial conditions will reliably produce the same results over multiple runs (to the level of precision being simulated). However, in the real physical world, if the parameters describing a system do not in fact *exist* in the world with infinite precision, as we have argued above, then such *ontological* imprecision will necessarily lead, in a system with chaotic dynamics, to a *genuine under-determination* of future trajectories by the present state of the system (Del Santo & Gisin, 2019; Ben-Yami, 2020; Gisin, 2021b). In such cases,

the evolution of the system will be exquisitely sensitive to details (e.g., the value of the one hundredth digit describing some macroscopic parameter) which are not only unknown to us, they are unknown to the universe itself—they are, by definition, ontologically fuzzy and undefined states. In one striking example of this, "gargantuan" simulations of a three-body system (three gravitationally interacting black holes) have demonstrated a sensitivity to the details of their initial conditions that is *below the Planck length*, the smallest possible subdivision of space. Such a physical system literally could not exist in a sufficiently defined initial state for its chaotic evolution to be 'actually' deterministic (Boekholt et al., 2020).

In fact, this may even be true in the solar system itself. Henri Poincaré famously recognised the "three-body *problem*", which results from the non-linear interactions between more than two gravitating bodies. The evolution of such systems (which would include our solar system) is, in general, highly sensitive to even slight variations in initial conditions. Because the mass of our sun dominates the gravitational forces in the solar system, this inherent indeterminacy in the evolution of these planetary orbits may take hundreds of millions of years to manifest, but it is still a real feature of their dynamics (Laskar, 1990, 2013; Boué et al., 2012). Newton himself recognised this, highlighting the import of "inconsiderable Irregularities…, which may have risen from the mutual Actions of Comets and Planets upon one another, and which will be apt to increase, till this System wants a Reformation." (Newton, 1730/1952, p.402)

Hence, it is quite possible for a physical system to be both chaotic *and* fundamentally indeterministic. And, indeed, the systems we are interested in—living organisms and the local world they inhabit—seem to be both. This means that the evolution of their physical states through time is genuinely under-determined not just at microscopic, but also *at macroscopic levels*. It is simply an intrinsic and irreducible feature of chaotic systems in our universe that their future is under-determined by their current physical state.

4.4.4. Determinism does not hold at any level

This survey of principles of quantum and classical physics shows that any strict version of the thesis of determinism—either universal *or* classical—rests on assumptions and mathematical idealisations that simply do not seem to hold in physical reality. As John Earman (2004) put it: "determinism succeeds only with a little – or a lot – of help from its friends" (p.12), where claims of determinism often amount to little more than "making a postulate of wishful thinking" (p.4).

Similarly, Barbara Drossel (2023) argues:

> "even though the supposedly fundamental theories are deterministic and time-reversible, these theories are in practice supplemented by irreversible and stochastic features. There is no reason to assume that irreversibility and stochasticity are only apparent. Such a claim is based on ideology and not on evidence." (pp.13-14)

Contrary to convention in the philosophy of free will then, physics has not proven that complete determinism holds *at any level*. Instead, the evidence from physics itself strongly suggests that strict determinism is false at both the quantum and the classical level. The current state of any system plus the laws of physics do not, in fact, specify a single future for all subsequent time points in infinite detail. Instead, they under-determine it, meaning genuine alternate possibilities exist and the future is radically open, from the perspective of physics. For some kinds of systems—like the orbits of the planets—the behaviour of the system at macroscopic levels will be *effectively* deterministic (such that, for example, the position of the centre of mass of the Earth within the solar system is highly predictable many centuries in advance to some finite degree of precision). But for chaotic systems—like atmospheric systems or nervous systems, or even the solar system considered over longer timeframes—their future states are genuinely *under-determined* by the laws of physics themselves. There is thus no good reason to accept classical determinism as our starting premise or focal point for discussions of free will, and many good reasons not to.

4.5. Determinism-Plus-Randomness

In **Section 4.5.1**, we examine the influence of the determinism-plus-randomness worldview within the philosophical literature on free will. In **Section 4.5.2**, we present arguments from physics against determinism-plus-randomness. And in **Section 4.5.3**, we propose an alternative model of indeterminism, which we call 'pervasive indefiniteness'.

4.5.1. The premise of determinism-plus-randomness in the free will debate

Where does this leave us? We have so far argued, in contrast to the literature's current conceptual terrain, and crucially *for reasons that are independent of any considerations about free will itself,* that philosophers of free will can and should abandon their interest in "the thesis that there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.3). If our arguments are right, then such determinism is simply not an accurate description of our macroscopic universe or of the universe as a whole. It is just not the case that everything in the universe evolves *necessarily* from one time-point to the next

(*contra* universal determinism), or that there exists indeterminacy at the lowest levels of reality but not at the higher, macroscopic levels (*contra* classical determinism).

For many, this conclusion may not be especially newsworthy (though we hope the details of the arguments presented above will help permanently exorcise Laplace's demon). That is because, as noted in the introduction (**Section 4.2**), many philosophers of free will are not officially committed to the truth of determinism, despite the structure of the theoretical landscape being as it is. Many acknowledge that our neural decision-making processes are not fully deterministic, and some (namely, libertarian philosophers of free will) even actively embrace indeterminism as an essential ingredient in their theories of free will.

However, in doing so, it seems to us that these philosophers typically rely on a picture of indeterminism that, to a surprisingly large extent, continues to posit the reality of (classical) determinism—a view which we have called 'determinism-plus-randomness'. On this view, randomness or indeterminacy is framed as something that gets *added* to an otherwise deterministic universe or an otherwise deterministic decision-making process. Consider, for example, libertarian philosopher David Palmer's definition of the libertarian view of free will as one which *requires* "our actions to be breaks in the deterministic causal chain" (Palmer, 2014, p.4). Or his colleague Robert Kane's claim that "indeterminism does not have to be involved in all acts done 'of our own free wills'" (Kane, 2019, p.147). What both of these conflicting statements share, we suggest, is the presupposition that determinism *remains* "Nature's default mode" (Earman, 2008, p.817), at least for macroscopic systems such as ourselves. For Palmer, determinism holds *except* in cases of freely willed action. For Kane, determinism holds *even* in the case of some (but not all) freely willed actions. The determinism-plus-randomness worldview therefore appears to differ from classical determinism with respect to only one additional claim: that deterministic goings-on *can sometimes* be disrupted, broken, or diverted by the occasional insertion of randomness into the relevant causal chains. (Hence why we categorised this worldview in the introduction as a version of 'near-determinism'.)

The visual metaphor that is often used to illustrate the determinism-plus-randomness concept of indeterminism is Borges' 'Garden of Forking Paths' (see **Figure 6**). In this, "the single line going back into the past is just that: a single line indicating 'same past'; while the multiple lines going into the future represent 'different possible futures'" (Kane, 2007, p.24; see also van Inwagen, 1990; Law, 2023). The points of bifurcation therefore represent occasions where randomness 'percolates up' to have an effect at the

macroscopic scale, creating diverging paths along which determinism (locally) holds until the next bifurcation point is reached.

While presumably not every libertarian endorses this perspective, a commitment to such a worldview is certainly evident in the ongoing debates among event-causal libertarian philosophers (and their critics) over *where* in the causal chain leading to an action indeterminism needs to occur in order for an action to be considered free (in the libertarian sense). Or, as Laura Ekstrom (1999, p.85) puts it: "Where precisely are the *gaps* that must exist in nature in the chain of deterministic causal links between events in order to allow for human free will?" This 'location question' is often taken to be one of the main problems that any viable libertarian account of free will must address (Franklin, 2013, 2018; Kane, 2016). And, indeed, one of the most common ways to group the different (event-causal) libertarian theories of free will is according to where exactly they *locate* the indeterminism (Franklin, 2018; Clarke et al., 2021), with so-called "deliberative" theories locating the relevant indeterministic bifurcation or 'branching point' in the early stages of a decision-making process (Ekstrom, 1999; Mele, 2006; Dennett, 2015) and "centred" accounts placing it at the moment of the decision or action itself (Kane, 1996, 2007, 2019; Balaguer, 2010; Franklin, 2018; Ekstrom, 2019).

Implicit in all of these disputes is an acceptance that, wherever the indeterminism is *not* located, determinism holds true. And wherever it *is* located, it depends on a random or probabilistic *intrusion* that disrupts the otherwise deterministic goings-on. But is this the right way to conceptualise indeterminism? If the arguments presented so far in this paper are right, then we suggest it is not.

4.5.2. Arguments against determinism-plus-randomness

The evidence surveyed so far gives us two good reasons to reject the 'determinism-plus-randomness' concept of indeterminism. First, as already mentioned, this worldview presupposes that macroscopic processes are 'by default' deterministic because it is tacitly assumed that classical determinism holds true in our universe. We have shown, however, that there is no good empirical reason to accept the truth of classical determinism and, in fact, several compelling reasons to reject it.

Second, the determinism-plus-randomness worldview seems to identify *individual* quantum events (e.g., wave function collapses) as the source of indeterminacy in neural decision-making. At least, this is what critics of libertarianism often take it to imply, as illustrated by comments like the following:

> "Indeterminism does not confer freedom on us: I would feel that my freedom was impaired if I thought that a quantum mechanical trigger in my brain might cause me to leap into the garden and eat a slug." (Smart, 2003, p.63)

Now, it is certainly true that isolated, particular quantum events *can* have influences on macroscopic systems. Obvious examples include the emission of a photon by an excited atom or the radioactive decay of an atomic nucleus, which are inherently random in both the time and direction of emission (Ginzburg & Syrovatskii, 1964). One need not conjure contrived scenarios like Schrödinger's cat to see how such events could have an impact at macroscopic levels. Cosmic rays from distant systems can, for example, alter electronic states in digital computers, leading to so-called soft errors in computer memory (Ziegler & Lanford, 1979; Baumann, 2005). Cosmic rays can also cause damage to DNA, leading to macroscopic effects such as cancer (Percival, 1991; Atri & Melott, 2014).

However, these scenarios refer to outcomes of singular random quantum events impinging on a system *from outside*. A more pressing question is whether individual quantum events occurring *within a system* can 'percolate up' to affect macroscopic processes, in the manner envisaged by determinism-plus-randomness. In the brain in particular, we can ask if an individual ion in some neuron 'zigs' instead of 'zags', could that alter the firing of that neuron at some crucial moment, in a way that could ultimately make the difference between someone deciding to do A rather than B? Such a picture is not generally supported by current evidence. First, while the processes of neural transmission and firing are noisy (Faisal et al., 2005; Deco et al., 2009; Rusakov et al., 2020), there is little evidence that they can generally be swayed by *single events* at the quantum level. Second, while decision-making processes are similarly noisy (Glimcher, 2005; Sanborn et al., 2025), there also is no evidence—in mammals at least—that they are generically sensitive to single firings of single neurons. On the contrary, neural systems are often designed to be largely insensitive to the precise details of *individual* neuronal firings (**Chapter 3**, published as Potter & Mitchell, 2025). Even Epicurus, who proposed that atoms must occasionally "swerve" to "loosen the treaties of fate" and thus open the possibility of free will, did not envisage the kind of <one swerve-one action> relationship implied by the determinism-plus-randomness worldview (Sedley, 1983).

Instead, the evidence surveyed above suggests that, rather than deriving from isolated random events at quantum levels, indeterminacy is in fact just a very general and pervasive feature of neural dynamics. On this view, we ought to understand indeterminism in the brain in terms of the *ubiquitous* probabilistic nature of microscopic goings-on, which *collectively* entail an under-determination of the macroscopic evolution of the system by

its current states. In a system where definite things are nevertheless happening, that fundamental indefiniteness necessarily gets resolved through interactions, apparently genuinely at random. The question then is whether and how this very general and pervasive process manifests at macroscopic levels.

As discussed in **Section 4.4.3**, the answer lies in the fact that many systems are chaotic, meaning that the evolution of their macroscopic states is inherently sensitive—*by default*—to the ways in which these low-level indeterminacies get resolved. We can see this, for example, in systems characterised by turbulence, where, even without factoring in any specific random *events*, the classical equations of fluid dynamics that describe these systems often do not have unique solutions (i.e., they are mathematically non-deterministic)—a phenomenon known as 'spontaneous stochasticity' (Eyink & Drivas, 2015; Neyrinck et al., 2022; Bandak et al., 2024). Indeed, macroscopic brain dynamics can and have been formally described as being 'turbulent' in direct analogy to such systems (Deco et al., 2009; Deco & Kringelbach, 2020; Deco et al., 2025), with neural networks often being poised at 'criticality', optimising stability for maintaining ongoing processes and flexibility for rapid adaptation (Hesse & Gross, 2014; Terada & Toyoizumi, 2024).

The upshot is that the kinds of systems most relevant to discussions of decision making and free will seem to be characterised by a *generic* noisiness of underlying processes (Deco et al., 2009; Tsimring, 2014) and a consequent openness of macroscopic evolution, rather than a sensitivity to *isolated* quantum events (*sensu* determinism-plus-randomness).

4.5.3. Pervasive Indefiniteness

The empirical arguments presented above thus support a far more radical concept of indeterminism that libertarians can make use of.[16] According to this conception, classical scales of reality are *themselves* indeterministic. First, in the sense that every classical system exists in a state that is, to some degree, indefinite: the states and properties of the system are pervasively 'fuzzy' insofar as that they lack complete ontological precision. And second, for non-linear macroscopic systems such as the brain specifically, this pervasive indefiniteness means that the way that the physical states of these systems will evolve over time is then *by default* under-determined by the low-level details of their current state. This means that 'Nature's default mode' for an agent's neural decision-making processes is simply *not* deterministic, as is supposed by a determinism-plus-randomness worldview.

[16] Moreover, we argue in **Chapter 5** that libertarians *should* make use of this alternative concept of indeterminism in order to fully overcome the infamous Luck Objection.

Under this new concept of indeterminism, it is therefore not appropriate to assume that decision-making processes would have "exactly one physically possible future", if they were somehow to unfold uninterrupted by isolatable intrusions of randomness. This is just an idealisation. On the contrary, for processes such as these, the '*default mode*' is an indeterministic time evolution—and, as such, our baseline expectation should be that, given the system's current state, at any given time-point, there will always be multiple physically possible ways it could evolve. To use Müller et al.'s (2019) terminology, our argument is that agents like us *inevitably* have a wide range of "real possibilities" available to us, where "what is *really possible* in a given situation is what can temporally evolve from that situation against the background of what the world is like" (p.3).

What this more radical conception of indeterminism means for free will philosophy is subtle but crucial. First, it means that mechanisms for the amplification of quantum effects need not be identified and cited in order to get off the ground the idea that the future of a macroscopic system is metaphysically open. This sort of openness comes for free in virtue of being a non-linear system in a universe where one's physical parameters necessarily contain *some* degree of ontological imprecision. Second, it means that indeterminacy should not be viewed as something that gets *added* intermittently to the evolution of macroscopic systems; it is not the result of an extra, *positive* sort of randomness that somehow intrudes into our universe to trouble the otherwise fully deterministic go of things, occasionally forcing upon us unwilled actions. It is, rather, a *negative*—a constant under-determination of the future by the low-level details of present states and the laws of physics, entailed by the unavoidable and pervasive indefiniteness (and thus openness) that characterises the present state of the system itself (Mariani & Torrengo, 2021).

One notable upshot of conceptualising indeterminism in this way is that the 'Garden of Forking Paths' is no longer the right metaphor for visualising the future of an indeterministic system. Instead, as Mitchell (2023a) describes, "[i]f we really could glimpse the future, we would see a world out of focus. Not separate paths already neatly laid out, waiting to be chosen—just a fuzzy, jittery picture that gets fuzzier and jitterier the further into the future you look" (p.161) (**Figure 6**).

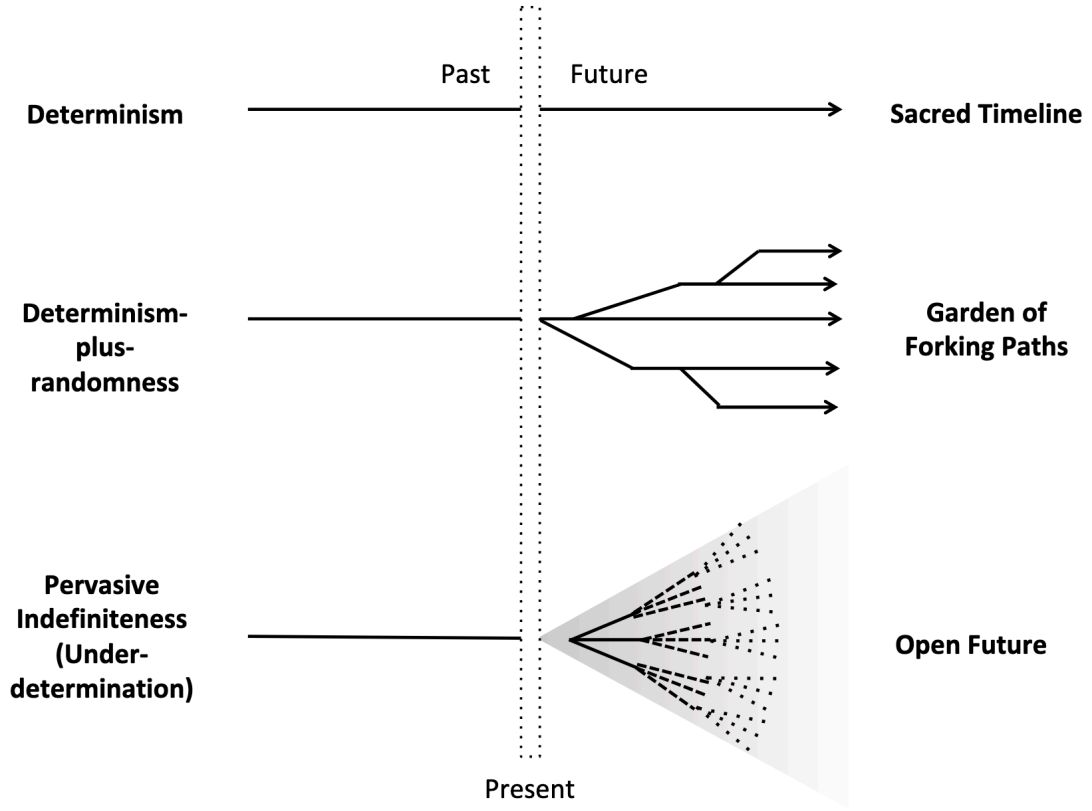

Figure 6 | Three metaphysical pictures of biological reality. ***Top***: Universal (Laplacean) determinism posits a single, unbranching timeline, with no *real possibilities*. ***Middle***: The standard way to conceptualise indeterminism ("determinism-plus-randomness") sees the future as a series of forking paths, where the alternate possibilities are in a sense pre-existing or pre-statable, and which branch is taken is determined by the intrusion of randomness at specific moments into an otherwise deterministic evolution. ***Bottom***: An alternative way to conceptualise indeterminism ("pervasive indefiniteness") which sees the future as radically open; with the near-future tightly constrained by current states, but the far-future highly under-determined beyond some time horizon. The present is then defined as the time during which events *occur*, resolving the open possibilities into actual happenings, which then become the fixed past (Smolin & Verde, 2021; Mitchell, 2023a).

4.6. Implications for the Free Will Debate

The arguments above have important implications for the framing of the free will debate. We contend that they call into question the relevance of many of the main arguments and dividing lines within this field. First, if determinism is false, then the existence question "*do we have free will?*" simply does not entail or rest on the traditional compatibility question "*is free will compatible with determinism?*", as it has traditionally been assumed to. It thus makes little sense to continue carving up the theoretical terrain primarily in terms of this question, as is currently the case where "the basic divide among philosophers is between compatibilism and incompatibilism" (Palmer, 2014, p.4).

Put another way, if our arguments above are correct, then there is simply no problem posed by causal determinism to the existence of free will, because the former does not hold. There is thus no need for philosophers to continue focusing so extensively on mounting attacks and defences of either free will itself or of moral responsibility against this non-existent threat. Instead, compatibilist philosophers in particular may be freed up to turn their attention toward a question that often gets gestured at in the literature, but is strikingly rarely given a concerted treatment: *are compatibilist accounts of free will compatible with indeterminism?* (but cf. Fischer, 2014; Sartorio, 2021; see also Mackie, 2018 for an in-depth discussion of the problem at hand).

Moreover, if we do not start with the defunct premise of determinism, then we no longer have to wonder: where does the 'freedom' required for free will come from? In an indeterministic universe, freedom (in the broad sense of *real possibilities* or *leeway*) comes for free. The more pressing question for a naturalistic philosophy of free will is instead: where does the control come from? How can agency emerge in the presence of such indeterminacy?

Crucially, answering this question depends on one's understanding of the nature of the indeterminacy itself—one's *concept* of indeterminism. Thus, the second major implication for the free will debate that emerges from the evidence and arguments surveyed above is the introduction of a new way of conceptualising indeterminism. Specifically, we have argued that randomness is not something 'added' to the world, something 'extra' that 'gives' you freedom. It is the noisy backdrop against which everything the organism does is set; a pervasive and irreducible feature of chaotic systems like us, and the challenge we are constantly facing. In **Chapter 5**, we show how this conceptual shift can help to flip the script on the Luck Objection that has traditionally plagued libertarian accounts of free will (Mele, 2006; Levy, 2011). In a world where the future is radically open, and a myriad possible things inevitably *could happen*, the organism must exert control to make happen *what it wants to happen*.

When seen through this lens, it becomes clear that the fundamental indeterminacy of low-level goings-on is the very thing that allows for macroscopic control—and thus organismal agency and free will—to emerge. Indeterminacy and randomness, *on their own*, are not sufficient for (or even overly conducive to) an agent's acting freely. Instead, what is key about the pervasive indefiniteness concept of indeterminism, as was recognised by Epicurus over two thousand years ago (Sedley, 1983), is that it means the low-level laws of physics are not "causally comprehensive" (*contra* Carroll, 2021). This enables the way systems are *organised*, at macroscopic scales, to also have some causal power in

determining how things go. Not by changing the underlying laws of physics, but simply by adding higher-order constraints (Tse, 2013; Ellis, 2016; Juarrero, 2023; **Chapter 3**, published as Potter & Mitchell, 2025). It is precisely this kind of "causal slack" that allows self-governing systems to emerge—ones that are capable of directing their own evolution at macroscopic scales, without having to control every microscopic element (Ellis, 2016; Ismael 2016; Mitchell, 2023a). In **Chapter 5** we show how such control can emerge and be exercised both diachronically, in the development of guiding reasons, and synchronically, in the execution of real-time decisions.

4.7. Possible Objections

We first look at some possible objections from physics to our arguments against determinism (**Section 4.7.1–4.7.3**), and then at a possible philosophical objection (**Section 4.7.4**).

4.7.1. Unitarity and the conservation of information

A common objection to the idea of fundamental indeterminacy in the evolution of quantum systems is that the Schrödinger equation is "unitary", meaning that information should be conserved as the equation follows the evolving system through time (Chiribella & Scandolo, 2015). This unitarity is taken to be such an essential feature of quantum theory that it is simply an unquestionable article of truth. And of course it does actually hold, but only right up to the moment when the system is 'observed', at which point the non-unitary collapse of the wave function occurs (Kastner, 2017), which absolutely does *not* conserve information (Drossel & Ellis, 2018).

Note that this view does not depend on an ontologically real interpretation of the wave function. We can treat it simply as a convenient mathematical object. The important point is that its characteristic unitarity does not apply to actual happenings. A similar feature emerges in the "division events" in the stochastic-quantum correspondence theory of Barandes, which interrupt the ongoing stochastic evolution of a system (Barandes, 2023, 2025). In either formulation, information is both gained and lost in such 'division' or 'observation' events. Gained in that the uncertainty of what will arise from a given probability distribution is resolved. And lost in the sense that that distribution cannot then be recovered from the resulting state of the system.

4.7.2. Time-reversibility

A related objection is that the equations of physics are time-reversible, which is again taken to mean that you cannot genuinely lose or create information as you go along. While it is true that *the equations* are time-invariant and symmetric, that does not mean the world is. As soon as people want to actually calculate or predict the behaviour of real physical systems, they have to supplement those equations with irreversible and stochastic features (Ellis & Drossel, 2020). The resolution of some systems therefore clearly involves time-asymmetric events, like division events or the collapse of the wave function.

4.7.3. The Block Universe

Einstein pictured the universe as a four-dimensional "block" of spacetime, with essentially no arrow of time (and no explanation for why it is always "now"). This supports a Laplacean position, where all of existence is simply given at once, thus reinforcing deterministic intuitions. There is, however, no need to accept this metaphysical interpretation, especially as it fails to explain why time seems to progress and why we occupy a particular time-slice that we call "the present". Many alternatives exist, which do entail genuine arrows of time (Layzer, 1975). These include, for example, the Evolving Block Universe (Ellis & Drossel, 2020), which expands both in space *and in time*. On this view, the time reversal invariance of the underlying equations is broken by the existence of a global arrow of time, established through the expansion of the universe through a hot early state to its present state. Moreover, it is a view that can explain the existence of 'the present' as the time in which the indefinite future becomes definite through physical interactions (or "events"), as in the theory of Smolin and Verde (2021).

4.7.4. Causal Determinism

Finally, some philosophers may object that we have been confusing the thesis of physical (pre-)determinism with the more general thesis of causal determinism. According to the former, the universe is such that 'there is at any instant exactly one physically possible future'. According to the latter, as described by Leibniz's principle of sufficient reason: "There is nothing without a reason, or no effect without a cause". An objector might contend that it is this latter notion of causal determinism that is at the heart of the free will debate and so our arguments have missed their mark.

Our response depends on what exactly the thesis of causal determinism is taken to imply. If it is taken to mean that there are no effects (events) that are not *necessitated* to occur by some antecedent cause (or set of causes), then we would argue that that just *is* a re-

statement of the thesis of physical (pre-)determinism (D'Ariano, 2018), and we hope to have shown why this is not an accurate description of our universe. If, on the other hand, the claims of causal determinism are meant to refer merely to the existence of cause-and-effect relationships in our universe (including probabilistic ones), then we would agree that our arguments do not speak against that. We are not claiming that human beings or other organisms are somehow exempt from cause-and-effect more generally. We instead follow Anscombe (1971) and others in rejecting the common underlying suggestion that "being caused implies being fixed in advance" (Runyan, 2024, p.112). Agents are indeed part of the causal nexus of the universe, but a part that can itself act as a cause.

Chapter 5

Chance, Choice, and Control: Free Will in an Indeterministic Universe

---

Status

Co-authored with Kevin Mitchell. This chapter is currently under review at the journal *Synthese*.

Author contributions

Equal contribution—both authors conceived, wrote, and edited the manuscript together.

5.1. Abstract

While the free will debate tends to focus primarily on the implications of determinism for freedom, a long line of philosophers have also argued that free will would not be compatible with *indeterminism* either. These arguments typically take the form of a so-called Luck Objection: a family of related arguments which all seek to show, roughly, that if an action is not causally pre-determined then it must be a sort of random happening, over which the agent lacks the control required for free will. If successful, these arguments are fatal for libertarian accounts of free will, which are committed to the view that free actions must be both undetermined *and* under the agent's control. In this paper, we defend libertarian free will against this challenge from luck. We argue that most formulations of the Luck Objection presuppose a conceptual model of indeterministic decision making that is not well aligned with recent advances in the natural sciences; specifically, we argue that they make assumptions about the nature of indeterminacy and about the causal structure of decision-making processes, which libertarians have good empirical reason (from both physics and neuroscience) to reject. We develop a more empirically plausible model of agential decision making and apply this to the problem of luck. We argue that, under such a model, it is entirely natural to think of an agent's actions as both 'undetermined' (in the sense of being under-determined) *and* under their own control. We conclude that indeterminism poses no threat to a more naturalistic version of libertarian free will.

5.2. Introduction

The "standard argument" against the existence of free will today is a logical argument consisting of two main parts (Doyle, 2011; see also van Inwagen, 1983). Either:

> Determinism—"the thesis that there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.2)—is true in our universe and agents therefore lack the freedom to ever choose or do differently than they in fact do.

Or:

> Determinism is *not* true in our universe. Indeterminacy and chance really exist and are a feature of agential decision-making and action selection processes. But this means that an agent's choices and actions are ultimately random and that agents therefore lack the freedom to control what they do.

In **Chapter 4**, we argued on empirical grounds, for reasons independent of any free will considerations, that determinism does *not* accurately describe the evolution of our

universe or of our neural decision-making processes. Instead, we showed that there are strong reasons to think that "Nature's default mode" is *in*determinism (Earman, 2008, p.817), of a sort that is particularly consequential for complex systems like us (see also Drossel, 2015; Ellis, 2016; Del Santo & Gisin, 2019; Gisin, 2021b). It is therefore our view that the second horn of the 'standard argument' is the relevant one for assessing the existence and nature of free will.

The second horn of the 'standard argument' expresses what is commonly known as a "Luck Objection". Proponents of the Luck Objection argue that "[i]ndeterminism appears to entail that it is not the *agent* who is the locus of control" over their own choices and action (Fischer, 2005, p.xxix, *original emphasis*). Instead, they claim, indeterminism introduces an element of chance or randomness into the agent's decision-making processes which serves to *prevent* the agent from having the sort of control required for free will, by making the outcome of these processes (i.e., the agent's eventual decisions and actions) a mere matter of chance—and, thus, a matter of *luck* for the agent herself. If successful, these arguments are fatal for libertarian accounts of free will, which are committed to the view that free actions must be both undetermined *and* under the agent's control.

Of course, proponents of the Luck Objection cannot just *assert* that agents would lack control over indeterministic decision-making processes and their "undetermined" outcomes. To do so would be to simply beg the question against the libertarian who, as Robert Kane (2019) puts it, already assumes that "there is a kind of control that indeterminism does preclude… *antecedent determining control* (ADC)—the power to guarantee or determine which of a set of outcomes is going to occur *before* it occurs" (p.155)—but is explicitly committed to the view that ADC *itself* is incompatible with free will. Instead, then, proponents of the Luck Objection must provide an argument for *why it is* that an action's not being causally necessitated to occur renders it a mere 'matter of luck' or *why it is* that an action being "undetermined" means that it is not under the agent's control (Steward, 2012).

There are therefore two key elements to a successful Luck Objection. First, one must posit (or adopt or just assume) a specific account for *how* metaphysical indeterminism introduces chance or randomness into the decision-making process. That is, one must select a specific *model* of indeterministic decision making to work with, either implicitly or explicitly. Then, one must provide an argument for why, under such a model, agents would lack the sort of control over their decisions or actions that is required for free will.

Different versions of the Luck Objection can therefore be distinguished according to the *model* of indeterministic decision making they presuppose (and, thus, the corresponding

argument for why such indeterminacy would preclude agential control). In this paper, we consider five of the most prominent versions, which we have termed: (i) The Reductive Luck Objection, (ii) The 'Objective Probabilities' Luck Objection, (iii) The 'Disappearing Agent' Luck Objection, (iv) The Contrastive Luck Objection, and (v) The Problem of Present Luck.

A question often left unexplored in these debates is whether the models of indeterministic decision making being presupposed by these arguments are, in fact, naturalistically plausible. For libertarians and other free will realists, this ought to be an important consideration, since a valid but demonstrably unsound argument poses no major threat to the existence of *our* free will in *our* indeterministic universe.

Our goal in this paper is therefore to defend free will against the Luck Objection by arguing that agential control *is* possible in the sort of indeterministic universe *we appear to be living in*.[17] In **Sections 5.3–5.5**, we consider three prominent versions of the Luck Objection—i.e., (i)-(iii)—and argue that each relies on a model of indeterministic decision making that there is good empirical reason to reject (**Table 2**). Our aim in doing this is twofold. Firstly, we want to show that, although these arguments may be logically valid, they do not necessarily pose a threat to the existence of *our* free will, or to the prospects of agential control *in the sort of indeterministic universe in which we appear to live*. Secondly, by making these arguments, it allows us to start to sketch what we take to be a more empirically plausible model of indeterministic decision making.

Luck Objection	**Counter**
Reductive luck objection	Macroscopic control
Objective probabilities	Real-time cognition
Disappearing agent	Holistic agent causation
Contrastive luck	Noisy agential doings
Present luck	Deliberative control

Table 2. Varieties of Luck Objection and the counterarguments offered here.

In **Sections 5.6** and **5.7**, we then take this model and evaluate it with respect to two additional versions of the Luck Objection, (iv) and (v). We argue that, while each of these

[17] In doing so, we complete our response to the 'standard argument against free will' (which was started in **Chapter 4**).

versions of the objection certainly gets something right about the *limits* of agential control under indeterminism, it does not follow from this that indeterministic agents therefore lack control over their decisions and actions *in toto*. Instead, the picture that emerges from the biology of indeterministic decision making is one in which agents have evolved to manage—and sometimes *use*—the noise and indeterminacy that they are inevitably faced with, to greater or lesser degree, depending on the demands of the decision-making scenario. Such agents cannot, do not, and need not exert control over every little micro-event involved in their decision-making processes. Instead, they generally can and do exert sufficient control to ensure that what they want to happen, *happens*—at the level of resolution they care about for obtaining their goals and staying alive. This, we suggest, just *is* the agent exercising the sort of control required for freely choosing what to do, even in the presence of indeterminism. Finally, in **Section 5.8**, we present an overview of decision making across diverse kinds of scenarios, and show how they involve a varying interplay of chance, choice, and control.

5.3. The Reductive Luck Objection

The first and most intuitive version of the Luck Objection is what we have called the Reductive Luck Objection. Proponents of the Reductive Luck Objection argue that agents would lack control over "undetermined" actions because what makes an action undetermined is that it is at least partly the result of some sort of random occurrence (e.g., a quantum event) in the agent's brain; and, by definition, agents do not have control over such random occurrences. Hence, indeterministic agents cannot exercise sufficient control in action to act freely (Moore, 2021).

This version of the argument is commonplace in more informal articulations of the Luck Objection, as demonstrated in the following examples:

> "If an atom in my brain suddenly veers off with a random swerve, it must do so "for no reason at all," and if this causes me to choose or decide something important, I am completely at the mercy of these random swerves." (Dennett, 2015, p.3)
>
> "Indeterminism does not confer freedom on us: I would feel that my freedom was impaired if I thought that a quantum mechanical trigger in my brain might cause me to leap into the garden and eat a slug." (Smart, 2003, p.63)
>
> "If the occurrence of a choice depends on the occurrence of some undetermined or chance events (e.g., quantum events) in the brain over which the agent lacks

> control, then whether or not the choice occurs would appear to be "just a matter of luck" rather than something the agent had control over and was responsible for." (Kane, 2019, p.154)

On the face of it, this reductive version of the Luck Objection certainly has some intuitive force. If an action is only "undetermined" because it is 'completely at the mercy of random swerves', then it seems valid to conclude that the agent would lack control over that action. How such an agent behaves would indeed be fixed by factors, such as quantum collapses, that are beyond the agent's control.

This argument, however, presupposes a very specific model of what it is that makes a decision-making process indeterministic—a model which, we believe, does not accurately describe *our* decision-making processes, as biological systems. On the model being assumed here, indeterminacy must enter the neural decision-making process via some sort of 'percolating up' process that amplifies specific random occurrences at micro (e.g., quantum) scales such that they have an effect on the *otherwise deterministic* processes unfolding at more macro (e.g., neural) scales. Free will skeptic Derk Pereboom (2022) summarises the model nicely when he suggests that:

> "For alternative possibilities [in decision making] to be significantly probable, there would *have to be* mechanisms that facilitate the "percolating up" of significant microlevel indeterminacies to the neural level, on the analogy of a Geiger counter that senses microlevel events and registers them at the level of the moving of a macrolevel indicator" (p.8, *our emphasis*)

In relying on this 'percolating up' model of neural indeterminacy, the Reductive Luck Objection is committed to a general conception of indeterminism that we have previously referred to as "determinism-plus-randomness" (**Chapter 4**). On this view, an indeterministic universe is considered to be one in which the "normal course of events" gets "peppered by indeterminism" (Smilansky, 2006; see also McKenna & Pereboom, 2016, p.16). That is, where indeterminism gets intermittently *added* to what is an *otherwise deterministic* universe in the form of occasional random events.

This determinism-plus-randomness worldview is arguably the most common way of conceptualising indeterminism within the free will literature, and it seems to lead naturally toward the conclusions of the Reductive Luck Objection. If an agent's decision-making processes are only indeterministic insofar as they are subject to occasional intrusions of randomness, then it *does* seem to be the case that the outcome of these processes (i.e., the agent's "undetermined" choices and actions) would be ultimately

dependent on, or indeed *caused by*, these random microscale events, in a way that would then make them effectively random and thus not under the agent's control in any meaningful sense.

Fortunately, however, there is no good reason to think that this model accurately describes *our* decision-making processes. In **Chapter 4**, we have argued that, from the perspective of physics, determinism-plus-randomness (and the 'percolating up' model it supports) is not the most plausible way to conceptualise indeterminacy in the brain. We will not repeat the arguments for this claim here, but, in brief, we showed that evidence from modern physics strongly undermines the idea that complete, universal, Laplacean determinism (where "there is at any instant exactly one physically possible future" (van Inwagen, 1983, p.2)) provides an accurate description of *our* universe, given quantum mechanics. The determinism-plus-randomness worldview accepts this but continues to hold onto an assumption that the macroscopic (or "classical") universe *is* still ordinarily deterministic. Indeed, it is precisely this assumption of classical determinism that *generates* the 'percolating up' view of neural indeterminacy in the first place; since, given this assumption, the only way for a choice or action to be undetermined *is* for its causal history to be sensitive to (and thus dependent on) random, amplified quantum events, over which the agent has no control.

We argue, however, that there is no good empirical reason to accept the assumption of classical determinism and, in fact, there are several compelling reasons to reject it (*especially* when it comes to nonlinear macroscopic systems like us). Consequently, it is entirely empirically plausible—and, we would suggest, *probable*—given the evidence from physics, that the evolution of macroscopic systems such as the brain just *is* inherently and intrinsically non-deterministic, *by default*.

We have called this alternative, more empirically plausible conception of indeterminism 'pervasive indefiniteness'. On this view, the implication is that the outcomes of decision-making processes are simply *under-determined* by the low-level laws of physics, not as the consequence of specific undetermined quantum events, but as their basic nature: the physical state of a person's brain at any given time, in conjunction with the low-level laws of physics, specifies *a possibility space* for how the system might evolve over time but nothing more.

The indeterminacy in this view is therefore not a *positive* force or an *added* element. It is just a *negative*—the future simply *lacks* definition, given the current state of one's brain (or of the universe as a whole). If this is right, then "Nature's default mode", for systems like us, is one in which the future is *always* radically open—where many things inevitably

could happen, at all times. Importantly, this does not mean that *just anything* could happen. Over short timeframes, this possibility space may be quite constrained (though not completely), with farther future states becoming progressively less defined (see **Figure 6**).

For our purposes here, the important upshot is that, while quantum collapses and random micro events certainly do still happen in a universe characterised by such pervasive indefiniteness, it does not make sense to identify these events as the *source* of an agent's "undetermined" action, as the Reductive Luck Objection does. The indeterminacy of decision making, on this view, is not traceable to specific random events that might '*cause* me to leap into the garden and eat a slug'. Indeterminacy is just a pervasive and ubiquitous feature of the processes themselves; it is the 'default mode' for systems like us.

More concretely, we might say (and we can *see*) that the neural systems employed in decision making operate with what are *fundamentally* noisy components (Faisal et al., 2005; Glimcher, 2005; Rusakov et al., 2020; Sanborn et al., 2025). Goings-on at molecular levels are inherently characterised by ubiquitous thermal fluctuations and probabilistic chemical interactions, which create a constant backdrop of noise within which the agent's decision-making processes must operate. If this is right, then the Reductive Luck Objection simply loses its intuitive force as an argument against *our* free will because, given the sort of systems *we* seem to be, it just would not be accurate or appropriate to identify the source of an undetermined choice or action with any specific (random) event over which we had no control. Hence, there is no reason to accept that agents lack control over their undetermined actions *on the grounds that* they are somehow being caused by random events.

We contend that the Reductive Luck Objection therefore does not carry the threat for *our* free will that its proponents think it does. One might reasonably worry, however, that the alternative, 'pervasive indefiniteness' model of indeterminism we have presented does not actually *help* the proponent of indeterministic free will in any way, either. If anything, it may seem even *less* likely that an agent could have control over their choices and actions when indeterminism manifests as this sort of *pervasive* and *ubiquitous* under-determination of the future, as opposed to some occasional bursts of randomness into what would be an otherwise reliable world of cause-and-effect. Does rejecting the 'percolating up' model of neural indeterminacy actually *help* the case for free will? We consider this question in the next section, before then returning to some more sophisticated versions of the Luck Objection.

5.3.1. The Emergence of Macroscopic Control

Part of the intuitive appeal of the Reductive Luck Objection rests on an implicit conception of causation—namely, that all the physical causes at play within a system come from the bottom-up, i.e., from microphysical interactions. For many, if these microscopic goings-on are deterministic, then it seems we cannot be in control of our actions (Kant, 1788/2002; James, 1884; van Inwagen, 1983; Kane, 1996, 2024; O'Connor, 2000; Steward, 2012; Pereboom, 2014). What the Reductive Luck Objection points out, however, is that, even if some of these goings-on are *in*deterministic, it still seems plausible that we would not be in control of our actions. That is because our actions would still ultimately be the product of impersonal laws of nature (deterministic, probabilistic, or a mixture of both), over which *we* have no control. Hence, what the Reductive Luck Objection reveals is that the apparent challenge to free will here does not come from determinism or indeterminism, *per se*, but from causal reductionism.

Given this, what is crucial about the pervasive indefiniteness concept of indeterminism, in contrast to determinism-plus-randomness, is that it is *not* committed to the idea that all of the physical causes within a system come from the bottom-up. On this view, the low-level laws of physics are neither fully deterministic *nor* causally comprehensive, as some have claimed (e.g., Carroll, 2021). That is, given the microphysical state of a system at some time, the laws of physics prescribe a 'default' possibility space, but it is still the case that within that space, many things could happen. This means that non-reductive causes are not, *a priori*, excluded (Kim, 1993; 2000)—i.e., there is some 'causal slack' within the system (Ellis, 2016). Other factors, such as the way a system is organised, can then *also* causally influence how the system evolves over time by constraining this 'default' possibility space further (Juarrero, 1999, 2023; Steward, 2012; Tse, 2013; Ellis, 2016; Farnsworth, 2018; Mitchell, 2023a; Jaeger, 2024; **Chapter 3**, published as Potter & Mitchell, 2025).

This is not invoking some kind of mysterious form of top-down causation. It is, in fact, utterly commonplace. For example, the flow of electrons in a computer is constrained by the hardware design, the software it is currently running, and what that software is currently doing (Ellis, 2008; Dasgupta, 2016). The electrons are still subject to the laws of physics, but the possibility spaces for the states of the system are constrained by these higher-order, macroscopic causes. Crucially, this is an example of constraint that is actively designed to *enable* some macroscopic functionality (Juarrero, 1999; 2023).

Living beings similarly exhibit macroscopic functionalities. Indeed, the most fundamental of these is simply persistence: living organisms do thermodynamic work to resist the

second law of thermodynamics and remain organised (Schrödinger, 1944/2012; Maturana & Varela, 1980; Deacon, 2011; Moreno & Mossio, 2015). In a world where many things *could* happen due to pervasive indefiniteness (most of which would involve their demise), living beings have to continually make *themselves* happen.

The important point for free will, then, is that the very existence of living beings is a demonstration of our ability to exercise macroscopic forms of control *in the presence of* ongoing indeterminism and noise. It reflects the fact that natural selection *must* be acting on such non-reductive forms of control. In the first place, this action is almost tautological: it is just that some macroscopic organisations constrain the 'default' possibility space in a way that makes them better at persisting than others and so are selected for. But with the invention of reproduction and the resultant processes of Darwinian evolution, selection for more specific adaptations could arise. In particular, these include systems for adaptively controlling behaviour in a dynamic and often hostile world. Organisms evolved systems—notably including nervous systems—to guide their behaviour in response to changing situations so as to best ensure their persistence (and that of their offspring). In effect, they evolved the ability *to do things for reasons* (Cisek, 2019; Mitchell, 2023a; **Chapter 3**), *in spite of* the noisy and indeterministic conditions they are continuously faced with.

Biology therefore seems to have found a way to capitalise on the causal slack inherent in the low-level laws of physics in order to 'build' agents with a form of control that genuinely inheres at the macroscopic level. Importantly, this agent-level control is neither necessitating nor random: it can never be exercised completely deterministically (to the lowest levels of resolution) due to the inherent noisiness of the components and systems involved, but it is also clearly not random as nature would not have been able to select for it otherwise. Instead, it seems to be a form of causal control that the philosophers Stephen Mumford and Rani Lill Anjum (2015) have termed 'causal dispositionalism'. On this analysis of causation, the appropriate modality for understanding how causes bring about their effects is neither necessity (i.e., "only one possible outcome" (p.2)) nor pure contingency (i.e., "anything could follow anything" (p.3)), but a *sui generis* dispositional modality in which "a cause can be thought of as something that tends toward an effect of a certain kind and often succeeds in producing it without there ever being any necessity that it would do so" (p.4). It is in this sense, we suggest, that biological agents have causal control over their actions despite the presence of indeterminacy and noise; they actively constrain their 'default' under-determined possibility space in a way that *tends toward* particular decisions and actions. Consequently, they very much *are* the "locus of control" of what they do.

This picture also implies (or seems to imply) some aspect of *choice*. It suggests that biological agents have options available to them to choose between, at least insofar as their actions are not pre-determined by their own physical state at some time-point, *t*, such that it remains physically possible that many things *could* happen at $t + n$. This capacity for *choosing* has, however, been challenged by some of the more sophisticated and subtle versions of the Luck Objection. These arguments diverge from the Reductive Luck Objection in that they do not suggest that indeterministic agents are simply being 'pushed around' by random events inside them. They accept that such agents *can* exercise a certain level of causal control over their actions in a way that allows them to 'act for reasons' (as we have been arguing in this section). What these arguments contest, however, is whether indeterministic agents are capable of controlling *which* (rational) choice or action they take, when they take it. In other words, the worry is not that these agents lack control *in toto*, it is that they lack the control required to *choose* how (or in what way) they will exercise that control; they lack the capacity to *choose* which of the physically possible futures becomes actual. This is sometimes referred to as plural voluntary control: "control over a set of more than one alternative courses of action: control over *which* of the open alternatives will become actual" (Schlosser, 2008, p.7; see also Kane, 2019, 2024).

We will now consider two versions of the Luck Objection that target *this* type of control specifically. Once again, we argue that these arguments rely on a model of indeterministic decision making that we have good empirical reason to reject.

5.4. The 'Objective Probabilities' Luck Objection

Many formulations of the Luck Objection rely on a notion of 'objective' or 'ground-floor' probabilities (van Inwagen, 2000, 2011; Mele, 2006; Shabo, 2011, 2020; Schlosser, 2014). On these versions of the argument, it is assumed that what makes a decision-making process indeterministic is that it has multiple physically possible outcomes, each with a prior "objective, ground-floor probability" of occurring that is less than one (van Inwagen, 2000, p.15). With this assumption in place, proponents of the 'Objective Probabilities' Luck Objection then argue that, since agents would necessarily lack control over these prior objective probabilities (they are, by definition, 'objective'), they would therefore also lack control over *which* (rational) decision or action results from this indeterministic decision-making process. That is, they argue that indeterministic agents would inevitably lack the sort of control required to choose *which* of the possible decision options available to them actually occurs. We call this the 'Objective Probabilities' Luck Objection.

The most prominent expression of this type of argument comes in the form of Peter van Inwagen's (2000; 2011) well-known Rollback Argument, which goes as follows. Consider an agent, Jane, who is deliberating about whether to holiday in Colorado or Hawaii next year. She realises she has good reasons for both options. She loves swimming in the sea and she has several books she's been eager to read for a while now. On the other hand, she loves hiking and is anxious to do something adventurous before she gets too old. After a long deliberation, she chooses, at time *t*, to holiday in Hawaii.

Now, imagine that immediately after Jane makes this decision, "God cause[s] the universe to revert to precisely its state one minute before [*t*]… and then let things 'go forward again'" (van Inwagen, 2000, p.15). Suppose God does this thousands of times. If Jane's decision is the result of an indeterministic decision-making process, and she has good reasons for both options then, van Inwagen suggests, "[a]s the number of 'replays' increases, we observers shall – almost certainly – observe the ratio of the outcome [Hawaii] to the outcome [Colorado] settling down to, converging on, some value." A value which van Inwagen calls the "objective, 'ground-floor' probability" of each decision's occurrence (*ibid*).

If this is right, then van Inwagen concludes:

> "If we have watched seven hundred and twenty-six replays, we shall be faced with the inescapable impression that what happens in the seven-hundred-and-twenty-seventh replay will be due simply to chance." (*ibid*)

The idea then is that, once we learn of the underlying objective probabilities that supposedly govern which destination Jane will choose, we will quickly become convinced of the idea that Jane's deliberations are nothing more than the 'playing out' of these ground-floor probabilities. Her eventual decision will appear to us to be little more than a random draw from a pre-established probability distribution—say, a 57% chance of choosing Hawaii and a 43% chance of choosing Colorado. As such, we will be forced to conclude that *which* specific course of action Jane takes, on any given "replay" (including the original one), is quite literally a matter of chance and, thus, not something that is actually under Jane's control.

Again, the 'objective probabilities' version of the Luck Objection certainly has some intuitive force to it. If indeterministic decision-making processes really *are* governed (or even just described) by prior "objective" probabilities in this sense, then it does become hard to see how Jane's decision to holiday in Hawaii on one replay (or her decision to holiday in Colorado on another) could ever be considered a decision that was *really* under

her control—at least insofar as the control in question is "control over *which* of the open alternatives will become actual" (Schlosser, 2008, p.7).

The standard libertarian response to this version of the Luck Objection has been to argue that, regardless of whether Jane chooses to vacation in Hawaii or in Colorado on a given "replay", in either case she will still be acting *for her own reasons* on *that* replay. She is therefore still exercising a form of rational control in how she acts, and that is enough—these libertarians say—to reject van Inwagen's conclusion that Jane's actions in the rollback scenario are 'due simply to chance' (Kane, 1996; Clarke, 2003; Balaguer, 2010; Franklin, 2018).

An alternative response, courtesy of Mele (2006), is to accept that it is *partly* a matter of chance that Jane makes the decision she does on any given replay. But insist that this does not mean that it is *completely* out of Jane's control which choice she will make, since it would have been Jane's own history of choices and actions that had ultimately shaped the set of objective probabilities that are now structuring her decision making in the present context.

While we believe that both of these responses certainly get something right, we think a more fundamental problem with the Rollback Argument is that it, again, presupposes a model of indeterministic decision making that seems naturalistically implausible, at least as a description of *our* deliberative processes. Specifically, the whole argument rests on an assumption that each of the possible outcomes of an indeterministic agent's decision-making process has a pre-statable, objective probability of occurring, *prior to the agent's engaging in deliberation*, which we see little reason to accept (for similar arguments, see Buchak, 2013; Lemos, 2021).[18]

This assumption of 'prior objective probabilities' is certainly not uncommon in the literature, either. Mele (2006), in fact, suggests something very similar when he likens indeterministic decision-making processes to the spinning of a tiny roulette wheel inside the agent's head. He writes—of an agent in a similar predicament to Jane—that "objective probabilities for the various decisions open to the agent are set... Larger probabilities get a correspondingly larger segment of a tiny indeterministic neural roulette wheel in the agent's head than do smaller probabilities. A tiny neural ball [then] bounces along the

[18] To be clear, advocates of the Rollback Argument often presuppose this causal model precisely *because* it is the model assumed by some of their libertarian targets. Mark Balaguer's (2010) account of libertarian free will, for example, leans extensively on the idea of there being underlying "moment-of-choice probabilities" for each of the decisions that are open to the agent.

wheel; its landing in a particular segment *is* the agent's making the corresponding decision" (pp.8-9).

We take there to be little to commend this view, however. Here, we argue that it is highly implausible that *our* decision-making processes can be characterised in terms of such prior objective probabilities and neural roulette wheels. We give two main reasons for this.

First, it is not at all clear what would even justify the claim that prior objective probabilities really *do* govern or describe our decision-making processes in the manner being assumed here. Such a claim would, of course, be justified in a deterministic universe, where Jane's choosing Hawaii would have an objective probability of 1 prior to her engaging in deliberation (indeed, prior to her birth). This claim may also be justified in a determinism-plus-randomness universe, where the outcomes of decision-making processes are arguably tethered to the outcomes of individual quantum wave function collapses, which may *themselves* be governed by objective probabilities specified in the Schrödinger equation.

However, once these two metaphysical positions are off the table (as we argue for extensively in **Chapter 4**), it becomes very difficult to see how one could motivate an assumption that 'undetermined' decision outcomes are necessarily matters of objective prior probability. It is, of course, trivial that after any number of "replays" there will always be *some* ratio of Hawaii-decisions to Colorado-decisions in these rollback scenarios (Müller, n.d., p.4). But there seems to be no substantive rationale for thinking that this speaks to some sort of 'true' or 'objective' probability distribution, which the observed ratio is "converging on". Instead, as Lara Buchak (2013) has argued, without any independent justification for this assumption, the Rollback Argument ends up bordering on circularity: it simply stipulates the very thing that it is attempting to *show*—namely, that the agent's choices and action are governed by prior objective probabilities. Skepticism about this version of the Luck Objection's core assumption therefore seems warranted on the grounds that we lack a metaphysical justification or grounding for it.

Second, not only is there no good reason to *accept* this assumption, there is also compelling reason to *reject* it (at least insofar as it applies to us, as biological systems). This is because, for the neural or psychological state of an agent *prior to deliberation* to precisely pre-specify the sort of objective, ground-floor probability distribution being assumed here, it would need to be the case that the agent's 'character' (e.g., her set of beliefs, desires, plans, values, and other dispositions) or the physical configuration of her brain, prior to deliberation, was somehow able to encode or entail a complete set of possible actions she might consider taking, for every possible context or scenario she

might encounter, *coupled with* the relative motivational weightings she has for each one's occurrence within that given context. Only *then* would we able to talk about the agent's psychological character or brain state, prior to deliberation, specifying a range of possible (rational) decision outcomes, each with a corresponding objective probability of occurring, *which the agent then has no ability to influence or change during the course of deliberation*.

However, when it comes to biological systems, and human beings in particular, such a situation is highly implausible. Humans have effectively infinite degrees of freedom in the actions we could take at any moment and the precise ways we could take them. We also encounter infinitely diverse scenarios and situations, which are often novel in their details. The idea that we could precisely pre-encode potential responses (as a set of pre-statable fixed probabilities) to every novel scenario therefore comes up against a massive combinatorial explosion that would seem to make the problem computationally intractable (Bossaerts et al., 2019; Rich et al., 2020). Evolution (and learning) can of course pre-wire some simple reflexes to isolated stimuli, and does so where such stimuli recur with enough regularity and where a consistent response is adaptive (see **Section 5.8**). But it does not seem physically possible that evolution could have pre-wired solutions to every scenario we encounter, in the manner that seems to be demanded by the 'Objective Probabilities' Luck Objection.

In summary, there seems to be no rationale or evidence to support the view that our decision-making processes can be characterised by prior objective probabilities, in any strong ontological (anti-Humean) sense that might strip us of control over our actions. We therefore suggest that, in assuming the existence of these prior objective probabilities, the 'Objective Probabilities' Luck Objection presupposes a model of indeterministic decision-making that does not plausibly describe *our* decision-making processes. And, as such, one need not consider this version of the Luck Objection a threat to the idea that *we*—as indeterministic, *biological* agents—have sufficient control to genuinely choose which of the available courses of action we take.

5.4.1. Real-time Cognition

What *does* happen during deliberation then? If the outcome of an organism's deliberative process is not causally pre-determined (*sensu* determinism), it is not 'completely at the mercy of random swerves' (*sensu* determinism-plus randomness), and it is not governed by 'prior objective probabilities', then what does explain it?

The answer, we suggest, is that what organisms like us have to do during deliberation is to actually *deliberate*. Evolution solved the combinatorial problem presented by the

complexity of our environments, not by attempting to pre-wire behaviours for every scenario we may encounter, but by inventing onboard control systems that support real-time deliberation and decision making. That is, it invented *cognition*. And cognition is what allows us to come to our own all-things-considered judgements about what to do in a novel scenario, *through the processes of deliberation* (Glimcher, 2003; Gigerenzer & Gaissmaier, 2011; Redish, 2013; Shadlen & Kiani, 2013; Ismael, 2016).

Agents equipped with the capacity for cognition can (and must) *do* work, during deliberation, to translate their pre-existing set of general dispositions and character traits into possible actions they could take within the specific context they find themselves in. They must then *do* work, in real-time, to figure out which of these possible actions to favour, by assigning relative weightings to the sets of reasons and interests underlying each option (McCall & Lowe, 2005, 2008; Lemos, 2021).

Indeed, this is precisely what the neurobiology of decision making suggests is going on during deliberation and action selection (Vervaeke et al., 2012; Jaeger et al., 2024; and see also Cisek & Kalaska, 2010; Redish, 2013; Shadlen & Kiani 2013; Mitchell, 2023a; Mitchell, 2025). Empirical data and theoretical models of decision-making processes in the neocortex and basal ganglia generally support a view in which complex decisions involve the gradual, noisy accumulation of evidence for *multiple different options* simultaneously (Gold & Shadlen, 2007; Ratcliff & McKoon, 2008; Steinemann et al., 2024). Mutual inhibition creates a dynamic competition between the options, which can be biased and influenced by ongoing cognitive operations elsewhere in the brain that feed into these comparative evidence accumulation processes. The process is then modelled as terminating only once the accumulated evidence for one of the options reaches a certain threshold (Gold & Shadlen, 2007), thereby implying an extended *period of deliberation* before any "commitment" to a decision can be reached (Thura et al., 2022; Yu et al., 2025).

Importantly, the parameters of the deliberative process are themselves flexibly tunable. The length of deliberation, for example, can be calibrated based on the uncertainty or urgency of the situation via mechanisms that adjust the height of the required evidence threshold (Cisek et al., 2009; Thura et al., 2022; Yu et al., 2025). Likewise, the structure of the process can be dynamically modulated, *during deliberation itself*, based on real-time computed tradeoffs between speed and accuracy (Chittka et al., 2009), and other factors reflecting the current decision-making strategy (Thura et al., 2022; Yu et al., 2025). The neural processes employed in decision making are therefore not only extended in time but also widely distributed across the brain (Steinmetz et al., 2019; O'Connell & Kelly, 2021; Khilkevich et al., 2024), modulated by state and behavioural context (McCormick et al.,

2020; Robson & Li, 2022), and organised in a nested hierarchy operating in parallel over different timescales (Ranti et al., 2015; Badre & Nee, 2018).

There are two key points to take from this. First, taken as a whole, the neurobiological basis of decision-making looks much more like a set of processes in which the agent dedicates time and resources to *actively figuring out* and *assigning* weights to different action possibilities, than one in which the agent is merely discovering a weighting or probability distribution that was somehow entailed by her prior set of beliefs and desires (or prior brain state). As the philosopher Robert Nozick (1981) put it, it seems that:

> "Reasons do not come with previously given precisely specified weights; the decision process is not one of discovering such precise weights but of assigning them. The process not only weighs reasons, it (also) weights them." (p.294; see also, McCall & Lowe, 2005; Lemos, 2021)

The second key point is that, given that these neural processes are inherently noisy and non-deterministic (see **Section 5.3**), it is undoubtedly the case that each process would go differently, *at some degree of resolution*, if one were to rewind the universe and let them play out again. There is therefore no reason to think that the way in which the system will come to 'weight' its reasons during these deliberations is likely to have some sort of pre-specifiable outcome or trajectory (Lemos, 2021). Importantly, though, this does not make these processes completely random either. They are all clearly being constrained by some sort of functional outcome or goal. And just because they are not being *perfectly* constrained—in the sense of following pre-determined or probabilistically pre-specifiable trajectories—it does not mean they are therefore out of the agent's control and at the mercy of chance. As we have already seen, a process can still be *controlled*, even if it is *noisily* controlled (Lemos, 2021). Causes do not need to necessitate their effects (Anscombe, 1971; Mumford & Anjum, 2015; Runyan, 2024).

One potential objection to this way of thinking comes from what we might call 'biological determinism'. Proponents of biological determinism typically grant that the precise outcomes of an agent's deliberative processes are never going to be perfectly pre-statable (as they would be if they were causally pre-determined or characterised by prior objective probabilities), and that agents must therefore *do* work during deliberation to arrive a rational decision or action, in the way we have been describing. However, they insist that the agent still ultimately lacks control over *which* (rational) outcome occurs. The reason for this is that (they claim) the psychological state of the system, prior to deliberation, still necessarily entails a specific 'optimal' or 'rational' outcome that the agent's decision-making processes will (or at least *should*) reliably head toward—even if, due to

indeterminacy in the process, there is no 'fact of the matter' as to when, how, or even *whether* they will reach this specified outcome. In other words, the claim is that the way that an agent will come to weight her different options during deliberation is pre-*destined*, even if it is not pre-*determined*; the agent is simply pre-disposed, prior to deliberation, to come to 'perceive' one of the available options as more 'optimally rational' or favourable than the others *during the course of her deliberations.* And, in this sense, she lacks the sort of control required to actively *choose*, in real-time, which (rational) decision or action to take because, in a way, her decision has already been made for her by her prior biological or psychological state (see Sapolsky, 2023 for an argument of this sort).

The first thing to note in response to this objection is that it is not a Luck Objection, since it is not an argument that the agent's control is being *precluded by indeterminacy* in the decision-making process. Insofar as it is an argument that the agent *lacks* control over their eventual decision, it is more of an argument against the compatibility of free will and determinism (not of free will and indeterminism, as is our focus in this paper).

Nonetheless, it is still worth noting there is little reason to consider this type of objection a threat to *our* ability to choose what we do. That is because its presupposition that, in any given scenario, there exists some sort of singular 'optimal' or even just 'exact' decision that the agent is pre-disposed to rationally favour is, in fact, strongly disputed by observations of bounded rationality in humans (Simon, 1990). It is well established that, in many scenarios, the expected utilities for the different actions we might take are genuinely indistinguishable, given the information we have at the time (and, possibly, *ever* could have) (Kahneman & Tversky, 1979; Simon, 1990; Glimcher, 2003; Glimcher & Rustichini, 2004; Gigerenzer & Gaissmaier, 2011). In such cases, there is simply no 'right' decision or 'exact' set of optimal weightings to be 'found' during deliberation, and thus no reason to think we are compelled to head toward any singular 'ideal' outcome given our psychological state and the information we have available to us. Instead, it seems far more plausible that agents like us are just doing the best we can, with the information we have, under the time constraints we face, with noisy, indeterministic components, to make our way through an under-determined future, in order to come to a decision that satisfactorily meets our interests and reasons—a process which Herbert Simon (1955) called "satisficing". If this is right, then we contend that that just *is* the agent exercising the sort of real-time "control over *which* of the open alternatives will become actual" that seems to be necessary for freely choosing what to do (Schlosser, 2008, p.7).

A lingering worry for this view, however, is whether it really is *the agent* that is exercising this control. Or whether such control is merely the product of complicated mental or

neurophysiological events and processes going on *within* the agent. This is the concern raised by the third version of the Luck Objection we will now consider.

5.5. The "Disappearing Agent" Luck Objection

A third version of the Luck Objection comes from Derk Pereboom's "Disappearing Agent" argument (Pereboom, 2014, 2017; see also Haji, 2004; Caruso, 2012, 2014). Though most formulations of this argument do typically include a 'prior objective probabilities' assumption, the force of the argument does not actually rest on this assumption in the same way that the Rollback Argument does (or any other formulation of the 'Objective Probabilities' Luck Objection). It is therefore possible to make sense of the 'Disappearing Agent' Luck Objection without accepting the existence of prior objective probabilities. As such, we will treat it as a separate version of the Luck Objection for the purposes of the present paper.

The disappearing agent argument is typically presented as follows:

> "Suppose that a decision is made in a deliberative context in which the agent's moral motivations favor deciding to A, her prudential motivations favor her deciding to not-A, and the strengths of these motivations are in equipoise. A and not-A are the options she is considering. The potentially causally relevant events thus render the occurrence of each of these decisions equiprobable. But then the potentially causally relevant events do not settle which decision occurs, that is, whether the decision to A or the decision to not-A occurs. Since, given event-causal libertarianism, only events are causally relevant, nothing settles which decision occurs. Thus it can't be the agent or anything about the agent that settles which decision occurs, and she therefore lacks the control required for [free will and] moral responsibility" (Pereboom, 2017, p.1)

If we set aside the idea that the agent's prior motivations here—the "potentially causally relevant events" for her decision—are somehow pre-specifying an (objective) "equiprobable" probability distribution for which decision will occur, what Pereboom seems to be suggesting is a Luck Objection of the following sort. He assumes (in accordance with the 'event-causal' views he is addressing) that what makes a deliberative process indeterministic is that the motivational "events" going on in the agent's brain, during these deliberations, simply leave it open as to which decision they will cause. If this is right, he argues, then agents would necessarily lack control over *which* (rational) choice or action results from an indeterministic decision-making process because they would

have no control over *which* of these prior neural or mental "events" ultimately ends up causing their decision. Instead, Pereboom says, "the agent 'disappears' at the crucial point in the production of the action" (2022, p.10).

The 'disappearing agent' version of the Luck Objection therefore also presupposes a very specific model of indeterministic decision-making. The argument is primarily designed to be a critique of event-causal libertarianism (e.g., Balaguer, 2010) and thus it first assumes an event-causal theory of action in which an agent's actions are said to be exclusively caused by particular events and states *within* it, such as particular beliefs and desires (or their neural realisers) (Davidson, 1963; also see Aguilar & Buckareff, 2010). To form a Luck Objection, this standard event-causal story is then combined with an additional assumption about how the presence of indeterminism means that these "antecedent conditions [simply] *leave it open* whether the decision in question will transpire or not" (Haji, 2004, p.144, *our emphasis*).

Here, we argue that, once again, there seems to be little reason to think that this model of indeterministic decision making accurately depicts *our* deliberative processes. There are two main reasons why.

The first concerns the argument's central claim that nothing about the agent prior to deciding—nothing about their motivations or internal dynamics right up until they make a decision—"*settles* which decision occurs" (Pereboom, 2017, p.1., *our emphasis*). Or, in Haji's (2004) more positive framing of the same idea: it is the claim that the agent's prior motivations and dynamics simply "leave it open" which decision will occur (p.144).

As Al Mele (2024a) has noted, there is an ambiguity as to what this claim could amount to. On the one hand, it could mean that nothing about the agent, prior to deciding, causally *determines* or *guarantees* the exact decision she will make—that is, it will always be 'left open', *to some degree*, due to indeterminism, when and how the agent will make her decision. On the other hand, this claim could be taken to mean that nothing about the agent, prior to deciding, causally *explains* or *accounts for* the fact she makes the decision she does—that is, the agent's prior motivations and reasoning, right up until the apparent moment of decision, simply do not offer any sort of *indication* as to which decision the agent might make.

Neither of these interpretations is viable, however. If the claim is interpreted in the former sense, then it clearly begs the question against libertarianism, for the reasons outlined in the introduction (**Section 5.2**) (i.e., by demanding *antecedent determining control*). But if the claim is interpreted in the latter sense, then it does not seem to accurately describe *our*

decision-making processes. That is because, as already discussed in **Section 5.4.1**, what characterises the deliberations of complex biological agents like us is that we *do* cognitive work, during deliberation, to *figure out* which of our prior motivations we favour (Lemos, 2021). It therefore may not be possible to explain or predict which decision an agent will make based on the strengths of her motivations *prior to deliberation*. But by the time she comes to actually make a decision (within these 'torn choice' contexts), she will have assigned relative weightings to these competing motivations which very much *do* explain and account for the decision she eventually makes (even if these weightings do not necessitate exactly when and how this favoured decision will occur). We therefore reject the assumption that, in indeterministic decision-making processes, it is simply 'left open' *which* decision will occur.

One might continue to worry, however, that even if there is an explicable rationale for why the agent decided in accordance with one set of motivations rather than another (such that it is *not* simply 'left open' what the agent will decide to do), it does still seem to be case that "the agent 'disappears' at the crucial point in the production of the action" (Pereboom, 2022, p.10). That is because, according to the event-causal framework under consideration here, an agent's decisions are said to be *exclusively caused by* a subset of motivational "events" *within* the agent. However, this invites the long-standing critique that, on this picture:

> "reasons cause an intention, and an intention causes bodily movements, but nobody – that is, no person – *does* anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part." (Velleman, 1992, p.461, *original emphasis*).

With regard to this aspect of the 'Disappearing Agent' Luck Objection, we in fact agree with Pereboom and other critics of the event-causal framework (e.g., O'Connor, 2000; Hornsby, 2004; Steward, 2012). It seems right to say that, if an agent's decision was simply an instantaneous "event" that was being exclusively and linearly caused by isolated events *within* it—i.e., by one set of motivations *or another*—then the agent *would* "disappear" from the causation of that decision. There would be no strong sense in which we could say the agent *does* anything in bringing about that outcome. Things just happen within it and then things happen in the world that involve it.[19]

[19] Note, again, that this aspect of the 'Disappearing Agent' argument is not actually a *Luck* Objection. This worry applies equally to *any* event-causal model of decision-making, whether it is indeterministic or deterministic (as is the case for most compatibilist accounts of free will). The concern here is that *causal reductionism* precludes agential control, not indeterminacy.

Fortunately, however, the neurobiology of decision-making gives us good reason to think that our deliberative processes are *not* reductively event-causal, in this sense of our decisions being somehow exclusively and linearly caused by a particular set of motivations or "events" within us.

5.5.1. Holistic Agent Causation

This therefore brings us to our second reason for thinking that the 'Disappearing Agent' Luck Objection does not accurately characterise *our* decision-making processes. There are two key aspects to this.

First, as mentioned in **Section 5.4.1**, deliberative processes in the brain (and, indeed, in phenomenology) appear to involve cognitive operations that *directly compare* competing motivations and reasons *against one another*, and then weight them accordingly. There is therefore no reason to think of decision making in terms of a singular, linear causal chain from one set of prior motivational "events" to a particular decision outcome (with an (unactualised) alternative linear causal chain from a competing set of motivations to a *different* decision outcome) (see, for example, van Inwagen 1983, pp.126-152; McCall & Lowe, 2008, p.746; Pereboom, 2014, p.34; Moore, 2023, p.1460) . Instead, in a sense, *all* of the agent's reasons and motivations seem to be involved in causally informing her eventual decision, that is, in making an all-things-considered judgement about what to do (Ismael, 2016). So, citing a particular set of reasons as *exclusively* causing a decision to occur does not seem to be the right way to think about *our* decision-making processes.

The second and more important point to make, however, is that it just does not seem appropriate to be thinking of our decision-making processes in terms of these event-causal, billiard ball-style chains of efficient causation *at all* (**Chapter 3**, published as Potter & Mitchell, 2025). On the model we have been presenting here, what a decision *is* is the whole system settling into a new dynamical state that *satisfices* all of the different forces, factors, and constraints operating within it. These constraints include all of the different reasons, weightings, and outcomes of other cognitive operations that have been carried out (often in parallel) *during deliberation*.

Crucially, this is a whole-system, holistically integrated process where the causation involved is not *actually* decomposable or localisable in the way that event-causal models generally assume (or as suggested by the idealised picture of the brain often used to study the neural mechanisms of decision-making) (Pessoa, 2022). Instead, we would suggest that the neurobiology of decision-making supports a more agent-causal perspective (**Chapter 2**, published as Potter & Mitchell, 2022): the agent-as-a-whole is *using* its prior

motivations and its ongoing reasoning processes to inform how it constrains the noisy goings-on within it (as best it can) so as to (non-deterministically) settle into a new global 'decision' state which satisfies its interests and the other outcomes of its cognitive operations (Genkin et al., 2025). In such a system, it is simply not right to think of a reason (or the "event" of having some reason) as directly *causing* a decision. Instead, the system-as-a-whole *draws* on (and actively weights) its different reasons in the process of *making* a decision (Potter & Mitchell, 2025).

On this agent-causal view, the agent is not merely the site of a sequence of internal events, it is an emergent control system—a dynamical structure that integrates information, constrains noise, and causes outcomes.[20] Such an agent therefore absolutely does not "disappear" from the decision-making process, nor does she cede the causation of her decisions entirely to events and states within her. She is causally involved all the way through the decision-making process. Thus, while in effect we *agree* with the 'Disappearing Agent' Luck Objection's critique of event-causal models of free will, we also suggest that there is good empirical reason to think that *our* decision-making processes are not reductively event-causal in this sense. We therefore contend that there is little reason to think that this version of the Luck Objection poses a threat to *our* ability to control which (rational) decision or action we take.

To summarise, we have so far argued that three of the most prominent versions of the Luck Objection—the Reductive Luck Objection, the 'Objective Probabilities' Luck Objection, and the 'Disappearing Agent' Luck Objection—do *not* successfully establish that human free will would be ruled out in the presence of indeterminism. We have shown that there is little empirical reason to accept the claim that *we* lack control over what we do on the grounds that either (i) our choices and actions are the direct result of individual random micro events, or (ii) that they are the outcome of what is effectively a random draw from a pre-established probability distribution, or (iii) that they are the product of indeterministic "events" within us that simply 'leave it open' which decision will occur.

In making these arguments, we have simultaneously outlined what we consider to be a more empirically plausible model for thinking about what *our* indeterministic decision-making processes look like. On this view, indeterministic agents, such as ourselves, exercise a (non-necessitating) form of macroscopic, agent-level control in bringing about our choices and actions which is (a) informed by (but not linearly or exclusively *caused* by) our reasons for action, (b) where those reasons are constructed, compared and weighted

[20] In the 'causal dispositionalism' sense of causation described in **Section 5.3.1** (Mumford & Anjum, 2015).

during deliberation in a way that would not be objectively 'pre-statable' prior to deliberation, and (c) where all of this is carried out against a backdrop of constant and pervasive indeterminacy which the agent must continuously overcome or navigate (as best it can, but only to the level it cares about) in order to make happen what it broadly wants to happen from a future that is under-determined by its current physical state.

The result, we suggest, is an agent who very much *is* the "locus of control" for her decisions and actions (*contra* Fischer, 2005, p.xxix). Of course, this will not be a complete, *antecedent determining* sort of control (meaning that the agent can precisely determine the action's occurrence *ahead* of time). But it *is*, nonetheless, a naturalised form of plural voluntary control: the agent, prior to deliberation, has multiple options available to her and she exercises control, during deliberation, on the basis of her prior reasons and her *real-time reasoning*, to select one of these options and bring it to fruition. In other words, she has "the ability to choose, for reasons, from (at least) two options" (Levy, 2011, pp.45-46) and exercises this ability by exerting "control over *which* of the open alternatives will become actual" (Schlosser, 2008, p.7).

In the remainder of this paper, we will consider two more versions of the Luck Objection whose presuppositions are (under certain readings) *compatible* with this proposed model of indeterministic decision making, but whose proponents insist that such agents would still lack control over *which* (rational) decision they make. We will argue that, while both of these arguments certainly capture something important about the *limits* of agential control and the role of noise in decision-making processes, they do not establish that agents therefore *lack* control in a way that would make their actions 'a mere matter of luck or chance' and would thus be a threat to free will.

5.6. The Contrastive Luck Objection

The fourth Luck Objection we are going to look at is what we have called the Contrastive Luck Objection, which has been advanced most prominently by Neil Levy (2008; 2011). This version of the objection argues, once again, that indeterministic agents would lack the sort of control required for free will because they would lack control over *which* of the (rational) decision outcomes available to them during deliberation becomes actual. In the Contrastive Luck Objection, this concern gets characterised in terms of the agent lacking a sort of control that would be required to rationally explain why the agent <chose A *rather than* B>. As Levy (2008) writes:

> "The agent's reasons explain why she is choosing between these options, and not others (Jane's vacation deliberations will not end with her deciding to vacation in Afghanistan, or to take up yoga). But they do not explain her choosing Hawaii over Colorado, or vice-versa. In other words, the [contrastive] luck objection focuses on the contrastive fact, < that the agent chooses to *Φ rather than to Ψ* >, even though she has strong reasons for both. What explains this contrastive fact? Invoking the agent's reasons do not help: they have, as it were, got her this far, but since (by hypothesis) she does not have decisive reasons for *Φ*-ing or for *Ψ*-ing, they can take her no further. Instead, the contrastive decision is left to chance. Nothing about the agent, her character, judgment or reasons, explains the contrastive fact; it seems that it is a matter of chance which option she chooses" (p.752)

The argument therefore is that if a choice or action is "undetermined" then we (supposedly) cannot give a reasons-based explanation for the 'contrastive facts' about that choice or action. In the case of Jane's vacation decision, for example, 'nothing about her character, judgement or reasons' is said to be able to explain the contrastive fact that she <chose Hawaii *rather than* Colorado>. Hence, it is inferred from this (apparent) lack of explanation that Jane must therefore lack control over *which* of these options she chooses. In this sense, the Contrastive Luck Objection argues for the same conclusion as the 'Objective Probabilities' Luck Objection and the 'Disappearing Agent' Luck Objection. But it does so by relying on an apparent *lack of explanation*, rather than by appealing to a specific causal model that seems to preclude the relevant form of agential control in some way.

For our purposes here, the important question is: does the Contrastive Luck Objection apply to the model of indeterministic decision-making we have been developing so far? Clearly, when characterised in the way Levy does in the preceding passage, it does not. That is because, as we have already explained in response to the 'Disappearing Agent' Luck Objection (**Section 5.5**), on the model we have been presenting here, it is simply not the case that there is no reasons-based explanation for why Jane <chooses Hawaii *rather than* Colorado>. On the contrary, there *is* an agent-involving, rational explanation of this contrastive fact about her decision. It is explained by the cognitive processes she engages in *during deliberation* to figure out which option to favour, by reflecting on and directly comparing the different sets of considerations and reasons underlying each option, and assigning them relative weightings accordingly.

What the Contrastive Luck Objection gets right is that not every aspect of Jane's deliberation will be perfectly predictable based on facts about her—her motivations, character, or reasons—prior to deliberation. The precise trajectory of her decision-making

process, and the exact way she comes to weight each of the available options during deliberation, is simply not pre-statable in this way (for the reasons surveyed in **Section 5.4**). Proponents of this version of the objection are therefore correct to say that 'nothing about the agent prior to *deliberation*' explains why Jane <chose Hawaii rather than Colorado>, since she did not have any decisive inclinations one way or the other *before deliberating*. (Indeed, this is just what *defines* the sort of 'torn choice' scenarios under consideration here: if Jane had clear reasons to favour one destination over the other, prior to deliberation, then she would not be 'torn' at all (See **Section 5.8**)).

However, it is *not* correct, on our model, to say that 'nothing about the agent prior to her actually committing to one course of action' explains or accounts for why she <chose Hawaii rather than Colorado>, because her deliberation precedes this commitment. Thus, Jane's (noisy) reasoning processes *during deliberation* and the relative weightings of her two available options *after deliberation* clearly do explain her choice. We can therefore reject the claim that the contrastive fact about Jane's decision is simply "left to chance" and that, for that reason, the decision is not under her control.[21]

5.6.1. Decision-Making is a Noisy Agential Process

In a universe characterised by pervasive indefiniteness, agents are simply not able to precisely determine exactly how things will unfold (see **Section 5.3.1**). Their causal control manifests in the form of a downward constraint on the 'default' possibility space for goings-on at the more micro-scale, *not* as a precise or necessitating form of efficient causation (**Chapter 3,** published as Potter & Mitchell, 2025).

For Jane, this means that she cannot have *complete* (contrastive) control over every aspect of the cognitive process through which she comes to assigns weights to her reasons. But this does not mean that it is entirely random and inexplicable why, by the end of deliberation, Jane weights her options in the way that she does. Again, this is explained by the (noisy) cognitive operations themselves: the work Jane has done to construct her options, accumulate evidence for them, simulate their potential outcomes, compare these simulations against one another, assess her confidence in these simulations and their comparisons, and meta-cognitively reflect on all of this (Fleming et al, 2012). Crucially, this fact will be missed if one does not recognise that these processes are exactly that: *processes* (McCall & Lowe, 2008; Steward, 2012; Müller & Briegel, 2018; Lemos, 2021). Conflicted agents engage in these continuous processes of cognitive deliberation, each step

[21] This rebuttal has also been forcefully made by John Lemos (2021) and Storrs McCall and E.J. Lowe (2005).

of which contributes to the ongoing comparative weighting of the reasons and interests underlying their considered options. The assigning of these weightings is therefore not just a one-shot, random occurrence or the product of "a punctuated series of chance events" (McCall & Lowe, 2008, p.746), which certainly *would* make it rationally inexplicable why an agent weights their reasons <one way *rather than* another>. Instead, it is the gradual outcome of an iterative, dynamic reasoning process, where variation in the process is inevitable due to pervasive indeterminacy and noise within the system, but where (as we have already seen) this does not make the process completely random. On this view, decision making is simply a noisy agential doing: a non-pre-statable, non-random process over which the agent exercises causal control, as best it can but only to the level required to satisfy its needs.

One additional point to note, courtesy of Robert Kane (2019; 2024), is that even though an agent does not exercise *complete* control over her deliberative processes, she can still articulate and defend the reasons behind the choice she makes and the considerations that led her there. This reflects a capacity to endorse and take ownership of outcomes even without *full* contrastive control over the process that generate them. Kane (1996; 2007; 2024) has described this as a sort of "value experiment" where we say, in effect, "I am opting for this pathway. It is not required by my past reasons but is consistent with my past reasons and is one branching pathway my life can now meaningfully take. Whether it is the right choice only time will tell. Meanwhile, I am willing to take responsibility for it one way or the other" (2024, p.98; see also Steward, 2012). Such agents therefore come to a decision *for their own reasons* and are able to actively endorse those reasons (*over* the potential alternatives) even if it is not *completely* (contrastively) rational how and why the agent came to favour those specific reasons.

Thus, while we think the Contrastive Luck Objection certainly highlights something important about the *limits* of an agent's control over their decision-making processes we see no reason to think this undermines the sort of agential control required for free will.

5.7. The Problem of Present Luck

The final Luck Objection we will look at is what is sometimes called the Problem of Present Luck (Mele, 2005, 2006; de Calleja, 2014; Clarke, 2019).[22] This version of the argument

[22] Mele (2005; 2006), it should be noted, has done more than most to formulate and articulate the Problem of Present Luck, but he does not himself endorse its conclusion.

homes in on a specific claim, common among libertarian philosophers, concerning what is supposedly the 'moment' of an agent's decision. As Al Mele (2024b) describes it:

> "Typical libertarians hold that an agent freely decided at *t* to A only if, given the past and the laws of nature, the agent was able right up to *t* to do something else intentionally at *t* than decide to A." (p.2392)

In the case of Jane's vacation decision, the "typical libertarian" would therefore presumably hold that, in order for Jane's decision to vacation in Hawaii at *t* to be a free one, it would have to be the case that, if we were to repeatedly roll back the universe a split-second to the moment immediately prior to *t*, Jane would at least sometimes intentionally make the decision at *t* to vacation in Colorado instead.

What the Problem of Present Luck points out, however, is that this would seem to require that Jane can (sometimes) intentionally make a decision that directly contradicts her own rational deliberations and judgement. It would mean that, on at least some 'replays', Jane is able to choose a different option to the one that she has come to favour and weight most heavily during her deliberations. This is because, in the given scenario, everything about Jane is imagined to be the same right up until *t*, *including* the trajectory of her reasoning processes and the relative weightings she assigns to the options she is considering. Yet it is claimed that she could *still* intentionally choose *either* of the available options *at time t*.

If this were true, though, there would seem to be nothing about Jane that could explain why she chooses in the way that she does; once again, her (contrastive) decision would be rationally inexplicable. Worse still, it would explicitly open the door to what Helen Steward (2012) has called a "kind of mad irruption" or "random upsurge of total irrationality" in Jane's psychological life (pp.169-170). For any decision, it would always be possible (indeed, for Mele's "typical libertarian", it is apparently *required*) that the agent might spontaneously choose contrary to their best rational judgement. That is, it "remain[s] possible that forces over which [the agent] has no control might intervene and prevent him from making the decision he really wants (indeed, longs) to make and, moreover, thinks it would be entirely sensible to make" (*ibid,* p.141). The result—according to proponents of the Problem of Present Luck—is that it would always be a matter of *good* luck for the agent if her actual world is one in which this "upsurge of irrationality" does not occur, and of course *bad* luck if it is. Consequently, it would seem that such agents can never *really* have control over their decisions and actions.

We consider this version of the Luck Objection to be successful in its critique of Mele's "typical libertarian". It decisively reveals that it is incoherent for libertarians to hold a view

in which, in order for an agent to *freely* choose A at *t*, they must have had it within their power, right up until *t,* to (intentionally) choose to do a substantively different action B at *t, contra* all of their reasoning in favour of doing A. In a physicalist universe, this sort of radical 'freedom to do otherwise' will inevitably come at the cost of rationality. It would therefore seem to *threaten* an agent's free will, not secure it. Thus, we suggest that the type of 'freedom' or 'agency' being targeted by the Problem of Present Luck is not the sort of freedom that *naturalistic* libertarians ought to be looking for.

Fortunately, under the account of freedom we have been developing in this paper, it is not the sort of freedom that *we* seem to possess, either; the radical 'freedom to (irrationally) do otherwise' is not a feature of (or concern for) the sorts of indeterministic agent that *we* seem to be. That is because, on our model, it does not follow from the fact that an agent is not causally necessitated, in the moments immediately prior to *t*, to <choose A at t> that they *could* therefore <choose B at t> instead. Rather, when an agent's deliberations move them decisively toward favouring option A, what this means, in dynamical systems terms, is that the system moves closer to settling into an attractor state that just *is* them 'committing' to action A. Within this model, it is not physically possible for the system to just instantaneously jump to an alternative attractor state that would constitute the system (irrationally) 'committing' to action B, instead. Such changes of mind *are* possible, but they take time and are the result of gradual shifts in how the agent is assessing and weighting the reasons underlying each of the options.

Thus, if we *are* the sorts of indeterministic agent we have been arguing for, then there is no reason for naturalistic libertarians to worry about the Problem of Present Luck or "upsurges of irrationality". This version of the Luck Objection is correct to point out that the radical 'freedom to do otherwise'—apparently coveted by many "typical libertarians"—is not, in fact, a freedom worth wanting. Fortunately, though, it does not seem to be a freedom that *we* have.

5.7.1. Yes, You Could Have Done Otherwise

On the model we have been presenting here, biological agents do not have a radical, contra-rational 'freedom to do otherwise', but it is still within their power to choose and do differently than they in fact do. In other words, even though agents do not have it within their power, in the moments immediately prior to *t*, to <choose action B ***at time t***> if their deliberations have led them strongly toward action A, it *is* still within their power to <choose B> ***at some future time-point***. That is because, on our view, an agent's cognitive control over her deliberative processes may lead her to constrain the possibility space of the *near*-future in a way that would preclude her from (irrationally) choosing B during this

time-period. But there is nothing stopping her from refraining from making a decision, continuing to cognitively deliberate, starting to change her mind (i.e., her relative weightings on her reasons), and *eventually* being in a position where it is again physically possible that she might make decision B instead. Agents therefore always have the power to 'change their minds' during deliberation, and even a very late ability to veto the actual execution of an action, if they discover good reasons to do so. However, like the preceding processes of deliberation, these 'changes of mind' are not *instantaneous* events occurring at some specific "time t"—they, too, are processes that *take time*.

This therefore speaks to a much more reasonable sense in which biological agents have the 'freedom to do otherwise'. On this view, it is (i) physically open to these agents to choose to do either A or B, *at the onset of deliberation*, (ii) under their control (i.e., 'up to' the agent) *which* of these options they come to favour *during deliberation*, and (iii) always open to them to (eventually) 'change their mind', right up until the moment of decision. We take it that this might provide a satisfactory compromise for many 'typical libertarians' that secures an 'ability to do otherwise', without falling foul of the Problem of Present Luck.

In sum, then, the Problem of Present Luck illustrates something important about the *limits* of agential control under indeterminism; namely, that indeterministic agents do not have the sort of radical, contra-rational ability to act *against* one's own rational deliberations, that many "typical libertarians" supposedly demand. We hope to have shown, however, that the sort of control agents *do* have, on the model of indeterministic decision-making we have been presenting, is still sufficient (and, in fact, preferable) for acting freely.

5.8. The Diversity of Decision-Making Scenarios

An underlying theme of our argument in this paper has been that the causal model of decision making one presupposes is central to both *how* one assesses the claim that indeterministic agents lack agential control, and what the implications of those assessments ought to be (i.e., are they relevant for judgements about *our* free will?).

However, there is no one-size-fits-all causal model for the many types of decision-making scenario we encounter as we go about the world. Instead, organisms, including human beings, employ different systems and strategies of behavioural control within different kinds of situations (**Figure 7**). In this paper, we have focused solely on conflicted or 'torn' decision scenarios (where the agent has good reasons for incompatible options and is highly invested in making a considered decision), as these are by far the most discussed types of decision in the free will literature. In these scenarios, we have argued that chance

and indeterminacy manifests as a sort of ever-present noise that the agent must navigate and constrain in exercising their rational deliberative control. But it would be a mistake to think we could generalise this account to other 'decision-making' scenarios, where the relevant causal models are demonstrably different, the degree and kind of agential control at play are different, and the role that chance and indeterminacy plays is therefore also different. Here we take a brief look at some of the other types of decision-making scenarios we did not discuss in this paper.

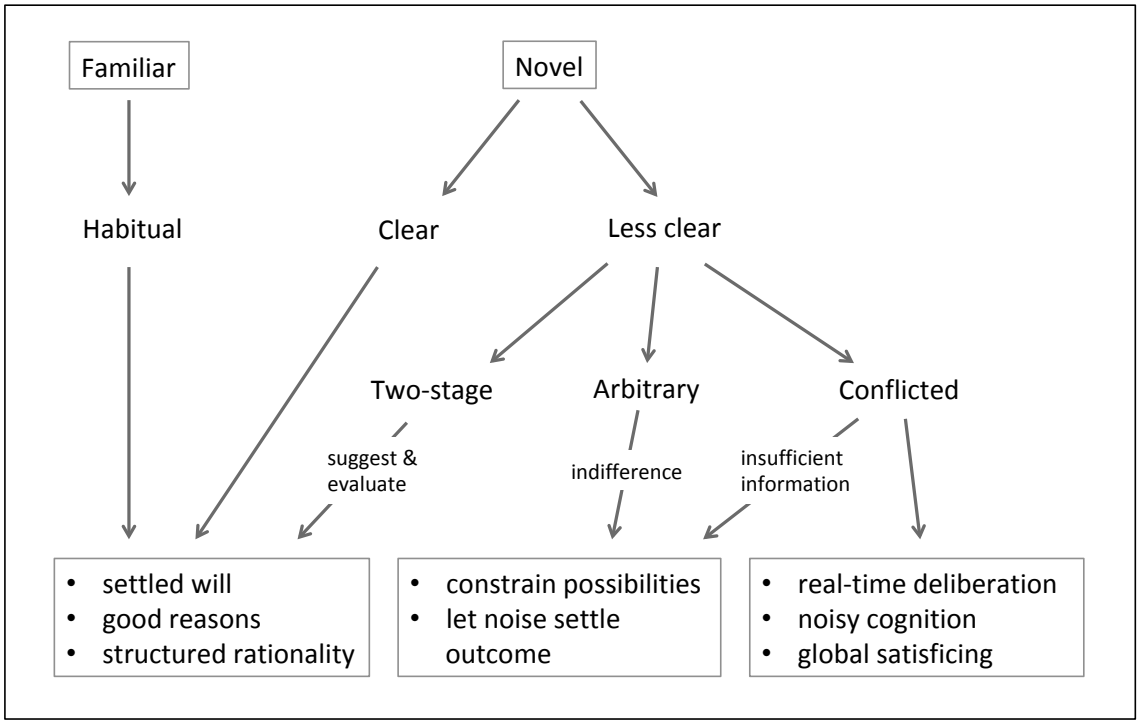

Figure 7 | The diversity of decision-making scenarios. See text for explanation.

First, as any experienced driver will know, in familiar scenarios much of our behaviour becomes habitual or automatic. This reflects prior learning from experiences we have had in similar scenarios, which has led us to cache some useful control policies (in addition to the ones that evolution may have pre-wired). In such scenarios, we do not have to think about what to do every time we re-encounter the relevant behavioural context, because we have already done that work in the past and we have paid attention to how those decisions turned out. Our habitual behaviour therefore issues from what we might call a "settled will". We are acting for our own decisive reasons, which we have accumulated through our own prior choices and actions (Mele, 2006). In this way, much of what we call "decision-making" might actually be thought of as a diachronically extended process of control by a temporally extended self (Mitchell, 2023a).

There is therefore no reason to take such behaviours as somehow undermining free will. It may very well be the case that in a rollback scenario for these kinds of habitual behaviours, we would do the same thing a thousand times out of a thousand trials. However, even if it is already 'settled' what we would want to do—i.e., we have decisive reasons that stand in every trial—it still seems reasonable to view this as an exercise of agential control, even though we are not actively *choosing* between options in real time. That is because, in a sense, we have already *done* the 'choosing' long before the "trial" began (Kane, 1999; 2024). This therefore reflects what has been called 'structural rationality', which rests more on psychological coherence than on active reasoning in the moment (Leffler, 2025).

A similar thing may be true in more novel scenarios, but where there is still one obviously preferred course of action. Helen Steward (2012), for example, presents a scenario where a man is deliberating about whether or not to move in with his girlfriend. All the parameters of the situation favour doing so, and align with all of his own prior motivations and values and aspirations. His choosing to do so would therefore, once again, be a reflection of a sort of "settled will". Just because he would always choose that option does not mean he is not in fact in control of that outcome.

What about novel scenarios where the best course of action is less clear? Again, these come in several varieties that call for different strategies. One of these was proposed by the psychologist William James (1884) and has come to be called the "two-stage model" of free will (Doyle, 2011). As we encounter a novel scenario we consider what our best options might be. This search will usually be informed by prior experience with similar *types of scenarios* in the past, which can be drawn on to suggest possible actions, given our current motivations. However, within this constrained search space, the ideas that actually *occur to us* are likely also influenced by noise in the dynamics of neuronal populations, specifically in the neocortex.

This is, however, just the first stage of the decision-making process. The second stage is to simulate the possible outcomes of the various options, weight the parameters according to what we deem most relevant, and thereby evaluate the possibilities and choose between them, releasing one action and continuing to inhibit all the others (see **Section 5.4.1**). This clearly entails evaluating the options relative to one's own reasons. James thus stated: "my thoughts come to me freely; my actions go from me wilfully". However, even the first stage reflects the learnings of the agent, the world model they have built up, and their own habits of thought that reflect their cached policies and accumulated reasons (i.e., a further manifestation of structured rationality (Leffler, 2025)). While there is some randomness at play, it is a constrained randomness.

Indeed, the degree of influence of this randomness is *tunable*. Under scenarios where an agent's goals are not being met by its selected actions, mechanisms may be employed to "go back to the drawing board" and widen the search space of possibilities. In vertebrates, such mechanisms include release of the neuromodulator norepinephrine from the terminals of cells in the locus coeruleus in the brainstem, which ramify throughout the neocortex. The effect of norepinephrine on cortical circuits is to "reset" the network, effectively allowing noise to shake the neurons out of previously favoured attractor states (representing the actions that had previously "sprung to mind"), allowing a wider exploration of other possibilities. This mechanism thus enables cognitive flexibility when past ideas no longer meet current demands. The important point is that the agent *can decide* to draw on the randomness in its own circuits to help them find a better line of behaviour.

There is another important class of scenarios where randomness can be exploited as a resource. These are situations where we do not know the best course of action or where we simply do not care very much about the outcome. In such situations, we may not have strong reasons to do any one thing in particular, but we usually have a good reason to *do something.* One option is to delay action until more information is available, or to act specifically to gather more information. This is a common behavioural strategy, observed across all kinds of animals, and is deployed at low frequency even in cases where we *do* have strong reasons to choose one behaviour. It thus not only helps when we do not know what to do, it also prevents over-commitment to a given course of action, when circumstances could change.

However, in many cases, no such information is forthcoming, and there may be no way to get it. In addition, we may not have the luxury of pursuing such an "infotactic" strategy—we often need to act more urgently. In such scenarios, we may allow the randomness in our neural circuits to break the deadlock and decisively bias the competition in favour of one option. This is a good strategy when we care about the outcome but don't have the resources to make an informed decision. But it is also a good strategy when we *don't care* about the outcome—when we are indifferent to what happens, because nothing is riding on it.

This likely applies in the famous "Libet experiments", which have provided so much fodder for philosophers of free will (Libet et al., 1982; Libet, 1999). The real deliberative decision in such experiments is taken when the subject agrees to take part and to follow the instructions of the experimenter. These instructions include moving their hand "whenever the urge strikes them" (i.e., explicitly for no reason) and reporting when they consciously

felt that urge. EEG recordings over the supplementary motor area reliably detect a ramping electrical potential (the 'readiness potential'), which begins several hundred milliseconds before the urge is felt, and which gradually approaches a threshold when the urge is felt and the action performed (or consciously vetoed). Further analyses have revealed, however, that: (i) this ramping up does not always happen—the potential can begin to increase and then go down again, indicating that the onset of the ramping does *not* in fact indicate the moment of commitment to an action (Schurger et al., 2012); and (ii) the readiness potential is not seen in conditions where the participant actually cares about the outcome (Maoz et al., 2019).

The best-supported interpretation of the results of the Libet experiments therefore seems to be that, because the agent has nothing at stake in when they make a hand movement, and because they have nevertheless agreed to make such movements occasionally, they allow circuits in the supplementary motor area that normally act as "evidence accumulators" (**see Section 5.4.1**) to instead accumulate noise until a threshold for action is met.

This: (i) undermines any sweeping generalisations from these experiments for free will in general, and (ii) illustrates a kind of mechanism that we probably employ quite regularly in our day-to-day lives. Many of our "micro-decisions" are inconsequential—the details of what we do or how we do it just do not matter much or at all. What is important is that we get on with things. We can thus *choose* to constrain our activities to the degree that we deem necessary and then *allow* the randomness in our neural circuits to play out and settle these inconsequential details.

The final scenario is one of decision-making under conflict, another favourite for philosophers' thought experiments. This is where we really *do care* about the outcome, but we have several viable, competing options open to us. These cases have been discussed at length above (**Sections 5.4–5.7**). To recap, in such scenarios, our processes of deliberation may be noisy, but the outcome of those processes will still reflect *our* decisions, made for our own reasons.

Across these different kinds of scenarios, we can thus see a varying interplay of chance, choice, and control in the exercise of agency. Indeed, organisms can even choose to exert control over the degree to which chance will influence their actions.

5.9. Discussion

In **Chapter 4,** we argued that the universe is not completely deterministic, at any level. The low-level laws of physics do not, in fact, prescribe a single future. Rather, the future is genuinely open, and many things can happen—particularly for non-linear, macroscopic systems like us. We suggested that we therefore do not need to frame the free will debate primarily around the thesis of determinism or focus so extensively on constructing arguments that defend free will or moral responsibility against this perceived (but likely non-existent) threat.

However, this still leaves a major challenge for libertarian views of free will—or for naturalistic accounts of agency more generally—to show how this kind of fundamental indeterminacy can *help* the case for free will rather than hinder it.

Here, we have attempted to meet that challenge. In particular, we have examined and addressed a number of Luck Objections widely believed to undermine libertarian positions. These objections can be cast in a new light if we: (i) reject the view of indeterminacy that we call "determinism-plus-randomness" and instead accept the picture of "pervasive indefiniteness" as a better match to the true physical ontology of the universe; (ii) recognise that causal control can be non-necessitating without being random (and thus *un*controlled); and (iii) recognise that deliberation is a process extended through time.

A key insight is that there is no reason to think that singular random events drive the actions of the agent. The "one swerve-one action" view (implied by the Reductive Luck Objection) maintains a position of causal reductionism that no longer holds (at least not necessarily) when there is some fundamental indeterminacy at play. Instead, this indeterminacy at microscopic levels introduces some causal slack and allows macroscopic structure to emerge, which can exert top-down constraint over the dynamics of the system.

Coupled with selection for persistence, and for macroscopic functionalities that favour it, this allowed the emergence of living beings endowed with behavioural control systems (Mitchell, 2023a). Such beings can act *for reasons*, whether these reasons are pre-configured by evolution or learned through individual experience. The question, from the perspective of free will, is how indeterminacy then figures into these processes of control and decision-making *in real time*. Specifically, does it mean that the actions of an organism are probabilistically pre-specified by all of its acquired reasons? Or that it is simply randomness and chance that settles the outcome of the decision-making process, thereby removing the agent from the equation?

We hope to have shown how the variety of Luck Objections that have been made to libertarian views can be surmounted or reinterpreted in the light of what we know about the neuroscience and psychology of decision-making and behavioural control. Organisms make decisions, for agent-level reasons, through processes of deliberation that extend through time. The resultant actions are not pre-determined nor pre-statable—many options are genuinely open for the agent to choose between *during deliberation*.

The agent's processes of decision making are holistic, integrative, and contextual, involving parallel processing in areas and circuits distributed across the brain. They are not instantaneous transitions from one state to another, but extended through time, *as the agent decides what to do*. This gets missed when one frames the problem space in terms of ballistic, irreversible 'moments of decision'.

Moreover, in this model, the agent is *not* passively driven by "events" happening within it, which generate further events, as envisioned in event-causal libertarianism. We argue instead for a naturalistic agent causation, which of course necessarily involves states and processes in the brain, but which cannot be wholly reduced to them (**Chapter 2**, published as Potter & Mitchell, 2022; **Chapter 3,** published as Potter & Mitchell, 2025).

Crucially, this kind of agent causation is neither mysteriously "contra-causal" nor absolute. To varying degrees and in different ways in different scenarios, the processes of decision-making may be influenced by the noisiness of the neural machinery that mediates them. The agent does not need to somehow 'reach down' and settle how individual random 'events' transpire. There is no way to and no need to micromanage the jitter of every molecule or the flow of every ion. Control is exercised at macroscopic levels, by broadly constraining the lower-level goings-on. In particular, control is exercised by setting and dynamically adjusting the criteria for neural firing, instantiated in the physical configuration of neural synapses and other parameters of neuronal physiology (Tse, 2013; Potter & Mitchell, 2025). These are the physical means that allow the organism to decide what to do by setting the relative weights for the parameters it deems to be relevant, through the processes of deliberation.

Agents thus operate under a Principle of Sufficient Control. This is manifest in two ways. First, they control macroscopic goings-on by constraining, rather than micro-managing, microscopic goings-on. And second, they exercise that constraint at macroscopic levels only to the degree they need to (and can). In situations where they have little reason to care about the outcome, they may allow the noisiness in neural circuits to settle the outcome. And in cases where they do care but where they don't have sufficient information to make an informed choice, they may similarly rely on stochastic processes to break the

symmetry. In both of these scenarios, a deliberative choice may be made to decide in this way. It is often more important *to do something* than to spend time dithering or hoping for more information. Control will thus be exercised in different ways in different kinds of scenarios, depending on uncertainty, urgency, salience, and other factors.

Finally, it is worth asking what the implications of this framework are for questions of moral responsibility. Such questions are, of course, often the prime motivating interest for asking about free will. Indeed, there is a common view that a freely willed action just *is* an action that one can be held morally responsible for. We have attempted to disentangle these issues, addressing here questions of agency, choice, and control more fundamentally.

The view we advance, of a naturalistic agent-causal libertarianism, may be used to ground notions of moral responsibility in a genuine "ability to do otherwise", though this is not absolute or unconstrained. It is neither the case that an agent's choices are fully pre-determined, nor that they are so undetermined as to be simply "lucky". Instead, we have provided what we hope is a more nuanced framework for thinking about decision making in terms of an interplay of chance, choice, and control. Importantly, these elements come into play in different ways across different kinds of scenarios, in real-time and through time. Recognising this fact about the *type* of agents we are, and the ways in which our control and freedom can systematically vary across these different circumstances, may therefore help to shine a new light on these old debates about moral responsibility.

Chapter 6

A Critique of the Agential Stance in Development and Evolution

---

Status

This chapter is an Accepted Manuscript of a book chapter previously published as:

> Potter, H. D., & Mitchell, K. J. (2024). A critique of the agential stance in development and evolution. In *The riddle of organismal agency* (pp. 131-149). Routledge. https://doi.org/10.4324/9781003413318

Author contributions

Equal contribution—both authors conceived, wrote, and edited the manuscript together.

6.1. Abstract

The claim that organisms are the agents of their own embryonic development, actively and purposively controlling and directing their own ontogenic trajectory toward adaptive outcomes, is central to an emerging set of heterodox perspectives within theoretical and philosophical biology. We refer to this view of development as the 'agential stance'. Several theoretical implications are claimed to follow from adopting the agential stance, each of which is taken to present a radical challenge to standard theories and approaches in evolutionary and developmental biology. In this chapter, we consider three of these proposed implications: (i) *Organism-Level Control*: the need for a causal framework for studying developmental biology that causally privileges the organism (as opposed to either privileging genes or embracing multi-level causation), (ii) *Agential Adaptation*: the need for an explanatory model of individual-level adaptedness that prioritises the developing organism's agency and purposiveness (as distinct from models that solely prioritise 'adaptation by natural selection'), and (iii) *Agential Evolution:* the need for an explanatory model of adaptive population change that prioritises the organism's agency and purposiveness in its development (as distinct from standard models of 'evolution by natural selection'). We argue here that if the agential stance is to be interpreted literally, in the sense required by each of these implications, then its claims are not well supported by the empirical evidence. If the view is to be interpreted heuristically, as its proponents sometimes suggest, then it cannot offer the sort of distinctive or novel explanatory insight needed to ground these radical implications.

6.2. Introduction

The agential perspective is an emerging conceptual framework within theoretical and philosophical biology, which seeks to foreground the view that organisms are agents and then explore the consequences of this insight for key topics in biology (Walsh, 2006, 2015, 2018; Sultan et al., 2022; Uller, 2022; Fábregas-Tejeda & Baedke, 2023; Nadolski & Moczek, 2023; Snell-Rood & Ehlman, 2023; Fulda, 2023; Walsh & Rupik, 2023; Jaeger, 2024). We have previously argued along these lines for the position that organisms *themselves* are the agents of their own behaviour (**Chapter 2**, published as Potter & Mitchell, 2022; **Chapter 3**, published as Potter & Mitchell, 2025; Mitchell, 2023a; see also Walsh, 2015; **Chapter 5**). On this view, organisms are not mere automata, driven around by complicated genetic or neuronal happenings, nor by the conditions of their immediate or historical environment. On the contrary, over ontogenesis and maturation they develop the capacity to *act* in the

world, with holistic, integrative, purposive, and goal-directed behaviours that are genuinely 'up to them' *qua* agents. Consequently, organismal behaviours are not generally amenable to a completely reductive analysis that abstracts away from the agent itself to find the 'real' causes of behaviour in neuronal, genetic, or atomic activity.

This agential view of behaviour has some important implications for how we conceptualise and understand evolution. First, it brings into focus the fact that what organisms within an ecosystem collectively *do* is fundamentally what shapes the evolutionary trajectories of lineages. Their actions alter or even primarily create the selective pressures and ecological opportunities to which the organisms within that niche adapt. In this sense, organisms are the entities that *enact* natural selection through their choices and actions (Walsh, 2015).

Second, and more particularly, the agential view identifies processes of niche construction and cultural inheritance as clear instances in which organisms play an active role in directing evolutionary change by purposefully modifying their own environments. Taking both these points, it becomes clear that "[a]daptive evolution does not unfold as populations migrate along fixed adaptive landscapes" within fixed environments (Nadolski & Moczek, 2023, p.12). Rather, the ongoing interplay between organisms and their environments, where individuals actively and continuously modify (and are modified by) their surroundings, exerts a large influence over how lineages end up evolving. The agency of behaving organisms is therefore an essential factor to include in any model of evolution—a view that should not be considered controversial.

More controversial is the view that an agential perspective is needed in the developmental domain, too. This is the argument that organisms are the agents of their own *development*, and that the exercise of this developmental form of organismal agency can actively shape the trajectories of evolution (Walsh, 2006, 2015, 2018; Sultan et al., 2022; Nadolski & Moczek, 2023; Snell-Rood & Ehlman, 2023; Fulda, 2023; Walsh & Rupik, 2023). We refer to this view as the 'agential stance on development' (sometimes 'agential stance' for short), in order to differentiate it from the wider agential perspective of which it is a part.

Our focus in this chapter is to offer a critical analysis of the agential stance on development and of its proposed implications for evolutionary theory.[23] We ask: what do the claims of the agential stance consist in? What is the evidence that is taken to support these claims?

[23] Our analysis is restricted specifically to claims about *organismal* agency, wherein developing organisms are argued to 'actively control' their own development by *directly* influencing their internal states. We therefore do not consider the role of cellular agency (e.g., Jaeger, 2024), developmental niche construction (e.g., Schwab et al., 2017), or general behavioural embryology (Gottlieb, 1976) during development.

Are they to be interpreted literally or merely metaphorically, as serving a heuristic function? What is taken to follow from adopting the agential stance on development? And are these implications valid, justified, or useful?

We conclude that if the agential stance on development is interpreted literally, then its core claims are not supported by the empirical evidence. If the agential stance is interpreted in a heuristic sense, which its proponents sometimes endorse, then it does not offer the sort of novel or distinctive explanatory insights it is commonly claimed to.

6.3. The Agential Stance on Development

6.3.1. Claims of the Agential Stance

The agential stance is a particular way of conceptualising biological development. Its central thesis is that "the development of phenotypes is under the active control of the developing organism" (Sultan et al., 2022, p.9). On this view, morphogenesis is not a passive process. Instead, it is posited to be the process by which developing organisms *actively* "direct their own development" toward particular outcomes, in a purposive and goal-seeking manner (Walsh, 2015, p.84), by dynamically controlling their own internal structure and function—much like "clay modelling itself" (Russell, 1924, p.61, taken from Baedke, 2021).

In other words, the agential stance is the view that development is an agential process; morphogenesis is something the embryo 'does', as opposed to something that merely 'happens' to it. As philosopher Denis Walsh, one of the leading proponents of the view, puts it, the organism *itself* is the "unit that exerts executive control over development", such that:

> "Proper development depends upon the capacity of organisms to assimilate, integrate and orchestrate the causal contributions from genes, epigenetic structures, tissues, organs, behaviour and the physical, ecological and cultural setting." (Walsh, 2015, p.157)

A number of radical conceptual and empirical implications are argued to follow from this view. For the purposes of the present paper, we have grouped these into three categories:

> (i) O*rganism-Level Control*: The agential stance is often characterised by the causal and theoretical privilege it ascribes to the organism, *as a whole*, in answering the question 'how is organismal form generated during development?'. In comparing the agential stance with other systems-focused approaches in developmental

biology, such as Developmental Systems Theory (Oyama, 2000), Nadolski and Moczek (2023) explicitly highlight organism-level control as one of the agential view's primary distinguishing features. While other approaches may recognise a multitude of causal factors and levels, the agential stance is unique in privileging the "ordering influence from the system as a whole" (p.5) and thereby positioning the embryo as the 'executive control' unit of its own development.

> "Insofar as any single entity can be said to 'control', 'regulate' or 'orchestrate' this widely distributed plexus of causes [in development], it is the organism as a whole." (Walsh, 2015, p.18)

(ii) *Agential Adaptation:* The second implication that is suggested to follow from adopting an agential stance relates to the problem of how to explain biological adaptation. Standard explanations of the fittedness of organisms to their environments tend to take the form of 'adaptation by natural selection'. According to this explanation, organisms that are better fitted to their environments (i.e., that possess advantageous traits) have a higher likelihood of surviving and reproducing, and thus passing on those advantageous traits to their offspring, where the phenotypic differences in question are at least partly attributable to heritable variants. Over time, the (combinations of) variants that are most likely to lead to the production of well-fitted organisms become more prevalent in the population, leading to the appearance of increasingly well-adapted morphologies.

Adopting the agential stance is suggested to offer an alternative or distinctive mode of explanation: the purposive agency of the developing organism. From this perspective, the observable adaptedness of developmental outcomes is taken to be the consequence of the organism's purposive 'pursuit of its goals' during development (i.e., its agency). Importantly, this is taken to be a complementary, but still *distinctive*, explanatory strategy for explaining organismal adaptation to that of adaptation by natural selection (Walsh, 2015; Fábregas-Tejeda & Baedke, 2023); one that is altogether more active and organism-centered, wherein "development is the manifestation of the purposiveness of organisms" (Walsh, 2015, p.162). The logic here is simple: if the production of organismal form during ontogeny is something the organism actively 'does' itself, not merely the passive or stereotyped consequence of (genetic) inheritance and natural selection (see (i)) then, by

extension, adaptation must to some extent be something the organism agentially 'does', too.

(iii) *Agential Evolution:* Building on this idea further, the third major theoretical consequence that is suggested to follow from adopting the agential stance lies in the reframing of adaptive population change as something that individual organisms, as *agents*, 'enact' through the control of their own development.

> "the agential perspective makes the following difference to the view on evolution. It explicitly represents evolution as the consequence of organisms' pursuit of their goals." (Walsh & Rupik, 2023, p.10)

Instead of conceptualizing evolutionary change as arising "genotype-first" from the passive processes of genetic mutation, inheritance, drift, and natural selection—with the progressive adaptation of lineages emerging as a statistical consequence—the theory posits that "adaptive evolution is caused by the adaptiveness of organismal development" (Walsh, 2015, p.158) which, in turn, is explained by the organism's purposive agency (see (ii)). This argument is explicit in Walsh's claim that:

> "By locating the cause of the adaptive bias in evolution in the adaptive activities of organisms, particularly in their development, the [agential stance] does not need to invoke natural selection to do the job." (*ibid*)

This view is thus congruent with the idea of "phenotype-first" evolution (West-Eberhard, 2003), where, in this version, the evolutionary trajectory of a species arises from adaptive developmental plasticity of individuals, which enables them to actively create novel forms that are better adapted to new environments. Genetics can then supposedly 'catch up' by selecting either pre-existing or new genetic variants that predispose to these new phenotypes.

In this chapter, we argue that none of these three implications is well supported. Our reasoning for this is two-fold. First, the core claims of the agential stance on development need to be interpreted literally in order for the conclusions drawn in (i)-(iii) to follow from them. Yet we argue that there is insufficient evidence to support a literal interpretation. Such a perspective might conceivably be supported if one focuses primarily or solely on *adaptive* developmental processes, but it dissipates given a broader sample of the developmental evidence base which sufficiently takes into account non-adaptive and maladaptive processes and outcomes.

Importantly, some proponents of the agential stance do expressly distance themselves from the literal interpretation of these claims, explicitly stating that there is an important sense in which the organism's 'pursuit of its goals' during development is to be understood non-causally (see Fulda 2023 for an extensive articulation of this position). We must confess we find such disavowals hard to square with much of the language and framing surrounding the agential stance (e.g., describing the embryo as 'actively controlling' and 'orchestrating' its own development). However, for our purposes in this chapter, it suffices to say, as our second line of argument, that if the claims of the agential stance *are* to be interpreted metaphorically or heuristically in this way, then the implications above simply do not follow from it—particularly (ii) and (iii).

Before turning to these arguments, let us first examine the claims of the agential stance in more detail. To do this, we break the approach down into its two constituent dimensions: a 'causal dimension' and a 'purposive dimension'.

6.3.2. The Causal Dimension: Holistic Causation

The causal dimension of the agential stance focuses on how to operationalise the idea that the organism can be a causal contributor to development *in its own right*. That means identifying the whole organism as a locus of causation of (at least some) developmental effects in a way that is not reducible to or derivable from the causal contributions of its component parts.

An important aspect of the agential stance's commitment to organism-level control is that it entails a so-called anti-'gene-centrism'. By demonstrating that the organism has control over its own development, proponents of this view explicitly challenge standard approaches in biology which are perceived to emphasise the genetic causes of phenotypic development. Thus, one of the position's main theoretical consequences is the proposed conceptual shift from "genes as the causal driver of development to agents constructing their own development" (Snell-Rood & Ehlman, 2023, p.4). On this model, development is an activity that the organism 'does', and not merely a passive consequence of its (genetic) inheritance.

A holistic, dynamical systems view of development is certainly appropriate. Rather than a simplistic view of a direct, isolatable relationship between specific genes and specific traits, it is well established that genes work in concert with one another, forming emergent gene regulatory networks that reciprocally constrain, influence and regulate the activity of their constituent genes. Gene regulatory networks therefore represent a system in which, *contra* reductionism, the dynamics of the parts (i.e., genes) are in fact dependent on the

dynamics of the whole; the activity of individual genes is influenced by the activity of the entire gene regulatory network of which each is an interactive part. At a level up, gene regulation, including epigenetic mechanisms, is sensitive to (and hence causally influenced by) the cellular system of which it is a part. We can then zoom out further to find cellular networks that dynamically constrain and influence the activities of individual cells. This continues until, ultimately, one is forced to recognise that this reciprocal, mutual dependence between parts and wholes occurs at all scales within the developing system—with 'wholes' at one scale appearing as 'parts' at another—right up to the level of the whole organism.

From this perspective, it becomes clear that a developing embryo or foetus is a deeply integrated and interconnected whole, with higher levels of organisation influencing lower levels in an entirely natural, non-mysterious way (Oyama, 2000; Jaeger, 2024; see also **Chapter 2**). This causal holism provides a natural mechanism through which developing organisms might causally contribute to their own development *in their own right*—as is necessary to motivate the causal dimension of an agential perspective.

However, causal holism, on its own, is not generally considered sufficient to attribute agency or 'executive control' to a system of study. This kind of loopy, reciprocally causal dynamic is, in fact, a fundamental feature of all complex systems, even non-living ones: when elements of any system become mutually entrained and interdependent, a dynamic emerges in which the activity of any individual part is influenced by the whole it is embedded in, whilst concurrently contributing to those global dynamics itself (Juarrero, 1999; Mossio & Moreno, 2015; Pessoa, 2022). In development, these holistic dynamics have already been recognised and codified in the form of Developmental Systems Theory (Oyama, 2000), without being taken as evidence of organismal agency (see Nadolski & Moczek, 2023). Instead, then, the agential stance also requires a purposive dimension—an argument to support the additional connotation that embryos influence their own development in real time 'with purpose'.

6.3.3. The Purposive Dimension: Goal-Directedness

In conventional usage, agency is not just irreducible, holistic causation. It also entails *purposiveness*, i.e., some sense in which the agent caused the effect *for a reason* or *in the pursuit of a goal* (Nadolski & Moczek, 2023; Potter & Mitchell, 2022, 2025; Mitchell, 2023a). Since goals cannot be free-floating, such a teleological framing requires the existence of a subject to which the goal or purpose is attributed in some sort of behaviour-informing capacity. In line with this, most articulations of the agential stance include a purposive dimension, which attempts to operationalise the idea that "the organism

purposefully molds itself" in ontogeny (Baedke, 2021, p.6, our emphasis) or that "development is the manifestation of the purposiveness of organisms" (Walsh, 2015, p.162; see also Nadolski & Moczek, 2023; Newman, 2023).

Support for this claim is generally derived from observations where developing organisms, in virtue of their tightly integrated architectures, respond to perturbations that occur during development with holistic re-adjustments to their dynamical regimes. For example, an environmental cue detected in a localised area of the embryo can trigger a seemingly co-ordinated whole-system response to that cue, while damage to a particular tissue or cell can trigger whole-system changes that activate compensatory mechanisms and pathways. These self-organising or self-regulating capacities of the organism are referred to in the agential stance literature as 'plasticity' (Walsh, 2006), capturing the notion of the responsive "mutual adjustment among variable parts in development" (Uller, 2022, p.11). This focuses on the real-time responses of the developmental system to individual perturbations (both internal and external) with an emphasis on these not just occurring within the organism or being carried out by the organism, but being *controlled by* the organism.

Sometimes these responses will involve buffering perturbations—such as, genetic mutations, molecular noise, and environmental disturbances, even including lesions that split the embryo in half—in ways that maintain functional stability. On longer timescales, this manifests as developmental robustness or canalization (Waddington, 1957; Wagner, 2013), and contributes to what biologist Ludwig von Bertalanffy (1969) called 'equifinality'—i.e., the propensity for systems to reach the same end state, despite different starting conditions and different developmental trajectories. Other times they involve adopting alternative dynamical regimes and developmental trajectories in response to particular environmental cues, in ways that favour organismal forms that are better fitted to those conditions. This manifests as what we call here phenotypic plasticity, referring to outcomes of *processes* of developmental plasticity that result in altered phenotypes, as opposed to those which result in robust attainment of typical phenotypes. These variations in outcome can be merely quantitative (e.g., smaller growth under restrictive nutrient conditions) or, in some striking cases, can manifest as qualitatively distinct phenotypes (e.g., temperature-dependent sex determination). In all cases, plastic self-organisation plays an essential role in enabling morphogenesis to reach an adaptive outcome.

The crucial point for the agential stance is the inference that the developing organism does not plastically readjust *randomly* or *passively* in the face of these perturbations. Rather, its

whole-system responses are, in some sense, directed toward adaptivity. It is for this reason that developmental plasticity is taken as evidence of the developing organism exerting a *goal-directed* or *purposive* influence over its own development. The types of plastic, holistic readjustments the embryo tends to undergo in response to genetic and environmental variation are often precisely those which enable it to robustly attain and maintain *its* 'goal' of persisting.

> "robust, plastic organisms produce the responses they do precisely because, under the circumstances, those responses are conducive to an organism's survival" (Walsh, 2015, p.202)

These two dimensions therefore jointly motivate the view that developing embryos really are the agents of their own development. The apparent adaptivity of developmental robustness and phenotypic plasticity, in virtue of being 'underwritten' by the embryo's self-regulating capacities, is posited to sufficiently establish the developing organism as actively and purposefully controlling and directing its own ontogenic trajectory toward well-adapted outcomes—and, thus, to constitute a naturalised account of the agential stance on development. In turn, this agential view of development, with the organism itself as the 'executive controller', is argued to motivate and justify the three theoretical implications of: (i) *organism-level control,* (ii) *agential adaptation*, and (iii) *agential evolution*.

Our aim in the remainder of this chapter is to contest these three implications, at least insofar as they are derived from an agential stance on development. Of particular note is the suggestion that this capacity to produce alternate phenotypic forms in a goal-directed, adaptive way is the primary source of evolutionary novelties:

> "Organismal plasticity, a manifestation of goal-directed purposiveness, underwrites much of the production of evolutionary novelty. These novelties are not adaptively neutral (Hu et al., 2020; Uller et al., 2020); they are adaptively biased. The goal-directed purposiveness of organisms appears to be essential to the explanation of the origin of evolutionary novelties." (Walsh & Rupik, 2023, p.12)

6.4. A Critique of the Literal Interpretation of the Agential Stance

With the arguments underlying the approach laid out more explicitly, we can now revisit the agential stance's core claims in order to highlight an area of possible ambiguity. The

claim that "development… is under the active control of the developing organism" (Sultan et al., 2022, p.9)—or that embryos purposively "direct their own development" in pursuit of their goals—can be interpreted in one of two ways. First, it could be seen not only as ascribing causal power to the embryo-as-a-whole, but as ascribing some sort of forward-looking or *directive* causal power; a means by which the organism can be viewed as *trying* to attain certain outcomes, even in a deflationary sense. We will call this the literal interpretation because we take it that this is generally what it means for a subject to actively direct a process toward an outcome, particularly when this is characterised as the subject 'pursuing its own *goal*'. On this reading, the two dimensions of the agential stance would be intimately combined. It is not merely that the developing organism exerts an identifiable influence over development; it is that it exerts *the* influence that explains why a particular outcome occurs (rather than another) in virtue of its own goals and purposes.

Importantly, many authors of the agential stance expressly state that "[a]scribing agency to a system in no way imputes to it intentions or desires" (Sultan et al., 2022, p.5); it "is not a claim that living systems must have cognitive representations and desires that guide their activity" (Nadolski & Moczek, 2023, p.2). While we agree that intentionality or cognition are not implied or required, we take it that these statements are still compatible with a minimal notion of directive causal power—in the manner implied by a literal reading of the claim that developing organisms are the "unit that exerts *executive control* over development" (Walsh, 2015, p.157, *our emphasis*).

Alternatively, as is sometimes suggested, the core claims of the agential stance could also be interpreted metaphorically or heuristically, in a sense that emphasises the important insight that organisms-as-a-whole exert a causal influence over development that is real and often necessary for attaining adaptive end states, but without committing oneself to the implication that it literally exerts some sort of *directive* causal influence, in real-time. On this reading, the stance's two dimensions come apart. Embryos causally contribute to the processes of development (via holistic causation) and these causal contributions are necessary to enable developmental processes that are amenable to a particular (teleological) mode of explanation that posits and cites the organism's 'goal-directedness' as its explanans. But crucially, the latter is not intended to be a *consequence* of the former. Instead, the organism's 'goals' are taken to explain its developmental trajectory and outcome in a distinctly non-causal manner (Fulda, 2023).

For reasons we discuss in the next section, we contend that only a literal interpretation of the agential stance can justify the implications outlined in (i)-(iii). However, as we will now argue, such an interpretation does not appear to be empirically well supported.

6.4.1. The Problem of Non-Adaptive Plasticity

The agential stance leans heavily on observations of *adaptive* directionality in development to justify its purposive dimension. Developmental plasticity is claimed to be a manifestation of organismal purposiveness precisely *because* it is seen as representing the organism's 'ability' to reliably direct its development toward adaptive outcomes (its 'goal'), via supple, holistic, real-time responses to changing conditions. On a literal reading of these claims, this implies that these whole-system, self-organising responses are in some sense inherently directed toward adaptivity, so as to justify being described as the organism's "goal-directed pursuit of its ways of life" (Walsh & Rupik, 2023, p.9). Yet, such a reading does not appear to fit with the empirical evidence.

Developmental plasticity is indeed a ubiquitous property of organismal development; it is an essential component of development's general robustness to noise, mutation, and environmental variance, without which species could not survive or evolve (Wagner, 2013). However, it is presumably uncontroversial to note that the capacity to robustly buffer or attune to the contingencies of development—and thus tend toward well-adapted outcomes—is not unlimited. Many environmental insults or genetic mutations *do* in fact cause the developing system to dynamically re-adjust in atypical and non-adaptive ways, often resulting in malformations of various tissues or organs, or the complete failure of development and death.

As the agential stance highlights, the robustness of developing systems prevents many genetic mutations from having deleterious phenotypic effects. But it is acknowledged by proponents of the view that this is only true until it is not. Many single mutations have very strong, sometimes devastating effects on organismal development. And specific combinations of these—or even just the overall burden of mutations—can also often push the developing organism toward a non-adaptive outcome. (Indeed, entire fields of biology and medicine are devoted to the study of situations where such robustness fails.)

It is less common in the agential stance literature to see it acknowledged that sometimes it is the processes of developmental plasticity *themselves* that cause the emergence of a pathological state. In many cases, the proximate effect of some perturbation may be fairly innocuous, but it may lead to a set of reactive, cascading effects that take the dynamics of the whole system out of its typical regime and into a pathological one. In some cases, this involves moving the dynamical system to an alternate, sometimes qualitatively novel, but maladaptive attractor state, one that has not been selected for but that simply arises as an unpredictable 'failure mode' of a dynamical system with complex, non-linear interdependencies. This is well studied, for example, in the case of epileptogenesis

(Neuberger et al., 2019; Lignani et al., 2020), and is also posited for many of the emergent symptoms of neurodevelopmental disorders (Mitchell, 2015; Durstewitz et al., 2021; Bartsch et al., 2023).

Thus, it is not only the case that the processes of developmental plasticity often fail to produce an adaptive outcome, they also can, under some circumstances, 'actively' *produce* a maladaptive outcome. This poses an important problem for the literal interpretation of the agential stance. What justifies the claim that adaptive developmental outcomes are the organism's 'goal', and that developmental plasticity represents its active, real-time *pursuit* of that goal, given these cases of maladaptive plasticity and pathological outcomes? An agential approach would presumably need to account for the latter as situations in which the agent *fails* to attain its goal, yet there appears to be little empirical support for this. If anything, evidence of the full range of developmental processes and phenotypes seems to contradict the stance's original hypothesis that plasticity is a purposive phenomenon and hence requires an agential explanation.

Similar concerns relate to the agential stance's interpretation of developing organisms' capacity to respond to environmental cues in adaptive ways, so as to produce alternative phenotypic outcomes that are well fitted to their respective contexts. A number of striking examples of such phenotypic plasticity, across a menagerie of strange and wonderful creatures, are often presented as evidence of the organism 'actively directing' its own development toward these adaptive end states. However, these case studies are noteworthy precisely *because* they are unusual. Most animals do not exhibit such plasticity with alternate, well-fitted phenes, and most environmentally sensitive phenotypic variation is not of this sort. On the contrary, in general, deviations from the "wild-type" phenotype of most animals are non-adaptive (Ghalambor et al., 2015). Again, this leaves the agential stance with some difficult conceptual problems, particularly if it is to be interpreted literally: why is adaptive phenotypic plasticity not a more widely observed phenomenon, if embryos in general are taken as having the power to actively and purposively pursue adaptive end states during development, including the production of "novel" phenotypes? And what justifies positing organisms as actively pursuing adaptive outcomes, in a literal sense, when most environmentally sensitive phenotypic variation does not fit this description?

For our purposes, these questions serve an important rhetorical function. They demonstrate how taking into consideration evidence of non-adaptive and maladaptive plasticity can effectively undermine the impression that the 'internal logic of development' (Alberch, 1989) is inherently and actively adaptive, in such a way that might demand

explanation in terms of the organism's real-time, agential pursuit of its goal to persist. When one focuses solely or primarily on cases of *adaptive* robustness and plasticity, and ignores or downplays those times when development is not able to buffer perturbations or when alternate phenotypic outcomes are actually maladaptive, it can conceivably create a false impression of the scale to which the self-organising and self-regulating dynamics of developmental plasticity are *necessarily* directed toward adaptivity (**Figure 8**). And, in turn, create the impression that these processes are a manifestation, or consequence, of the "supple goal-directed, compensatory capacities" *of the organism* (Walsh, 2006, p.773; see also Walsh & Rupik, 2023)—akin to a real-time "decision-making process" (Snell-Rood & Ehlman, 2023, p.5). But, as we have seen, this is not a fair reflection of the full developmental evidence base.

As such, there would appear to be no principled reason to take developmental plasticity as *evidence* of 'organismal purposiveness' or of the organism 'actively pursuing its goal' of adaptation. At least, not if these claims are to be interpreted in the literal sense demanded by the assumption that *(ii) agential adaptation* and *(iii) agential evolution* follow from adopting the agential stance. These purported implications rely on the organism's agency (i.e., its purposiveness or goal-directedness) being able to offer a novel or distinctive explanation for *why* individual cases of plasticity—and development, more generally—turn out adaptive (rather than non-adaptive or maladaptive):

> "Organismal purposiveness *underlies* the contribution of development to adaptive evolution" (Walsh, 2015, p.159, *our emphasis*)

However, it appears as though 'organismal purposiveness' is just another way of *describing* instances where development robustly tends toward an adaptive outcome. We see no reason to think it 'underlies' or 'causes' that adaptivity, in a manner that could support a literal interpretation of the theory. And, thus, we see no distinctive role for this developmental form of agency in helping us explain adaptive population change.

We therefore suggest that a literal interpretation of the agential stance's claims is untenable, or at least unsubstantiated, and only *appears* viable under a highly selective reading of the developmental evidence base.

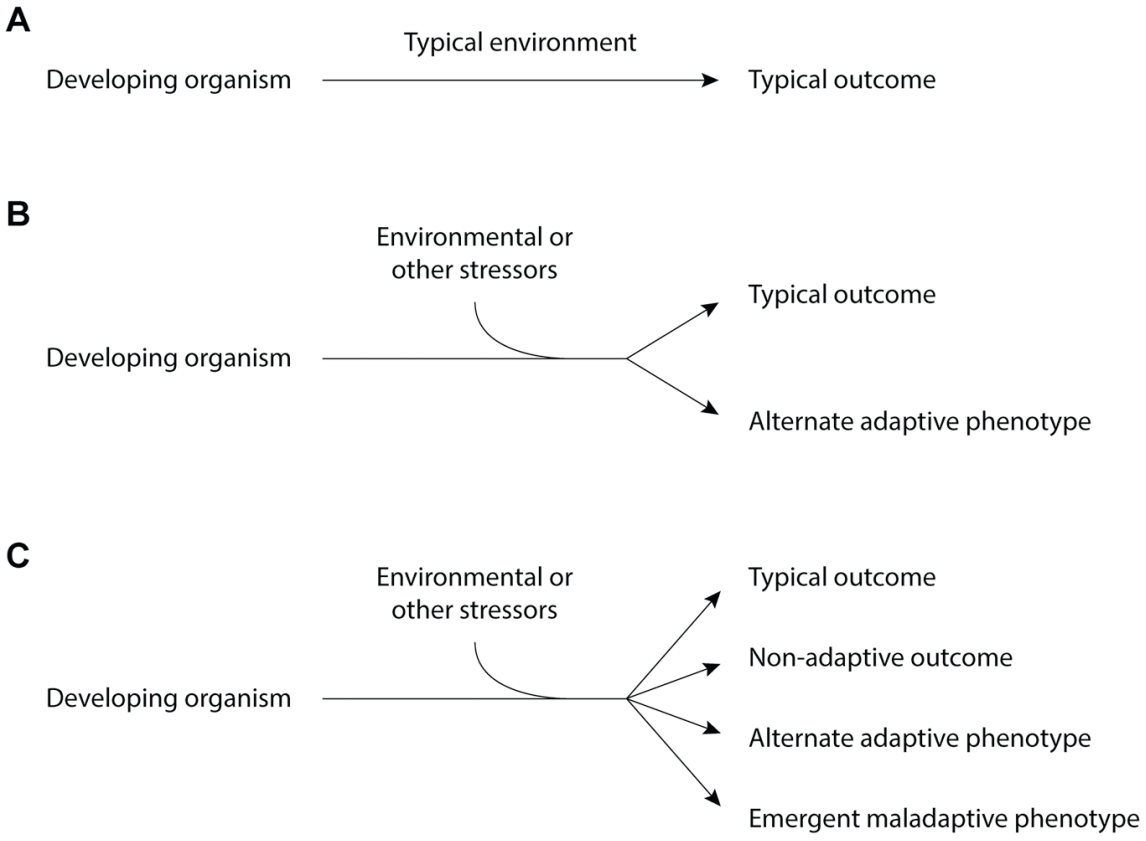

Figure 8 | Alternate views of phenotypic variation. (A) The observation that organisms develop toward species-typical outcomes under species-typical environmental conditions is congruent with the standard view of development as a passive 'happening,' the necessary 'playing out' of interactions between inherited components, with no 'subject' in charge. (B) The observation that developmental plasticity, under atypical environmental conditions or the presence of other stressors (including genetic variants), results in either robust attainment of a typical outcome or adaptive phenotypic plasticity implies an apparent purposiveness, which is taken to support a novel perspective in which the organism *actively* controls its development in a goal-directed manner. (C) A more even-handed survey of the evidence, which includes non-adaptive outcomes and 'actively' maladaptive ones, deflates this impression of purposiveness and leaves a more neutral view of developmental robustness and plasticity as evolved, holistic but passive tendencies of developmental systems that do not entail any sort of real-time, executive control.

6.5. A Critique of the Heuristic Interpretation of the Agential Stance

Some proponents of the agential stance pre-empt these challenges to the literal interpretation of their claims, by articulating a more heuristic version of the theory. On this view, the value of invoking concepts of agency, purposiveness and goals in development is said to lie, not in its identification of a literal, directive causal power exerted by the

organism, but in elucidating a teleological (or, rather, teleonomic) aspect of organismal development. By appealing to the embryo's purposiveness, it is argued that we obtain a previously unexplored explanatory lens for understanding development, one which can answer *why* a particular developmental process or outcome occurred and not just *how* it occurred (Walsh, 2015; Fulda, 2023).

Importantly, on this model, the goal that is imputed to the organism—which provides the answer to 'why' a developmental event occurs—does not help to *cause* the event in question. Instead, goals and events stand in a non-causal relation to one another called "hypothetical necessity", wherein "without the action [or event] in question, the goal would not have occurred and, with it, the goal… occurs reliably." (Walsh, 2018, p.173). Here, goals 'hypothetically necessitate' developmental events insofar as, if the organism has a particular goal, it necessitates (non-causally) that specific developmental events will occur. Goals are therefore taken to provide non-causal *explanations* of events using an analogous logical structure to causal explanations, and hence are argued to represent a scientifically valid form of explanation (see Walsh, 2015, ch.9).

Accepting this alternative teleological (or agential) form of explanation is then said to equip us with the explanatory apparatus to explain *why* developmental events and processes occur in the way they do (rather than another possible way), i.e., embryos respond to their conditions in precisely the way that *conduces* to their apparent goal of adaptation. A crucial supposition is that this sort of explanatory power is missing from existing evolutionary models. Hence, a (heuristic) agential stance on development is justified and necessary precisely *because* it offers a distinctive and ineliminable means of explaining individual-level adaptation (see (*ii) agential adaptation*) and population-level adaptation (see *(iii) agential evolution*). In essence, developmental and evolutionary processes can be seen as tending toward morphological adaptation because individual organisms are (hypothetically) 'pursuing their goals' during development, sometimes producing novel, adaptive phenotypes in the process:

> "Novel phenotypes…occur when they do *precisely because* they contribute to the organism's goals of survival and reproduction. These are not chance occurrences. We need to invoke the capacity of organisms to pursue goals in order to explain the origin of adaptive novelties" (Walsh, 2015, p.203)

We suggest that this heuristic interpretation of the agential stance faces the same set of questions and problems discussed in **Section 6.4.1** if it is taken to apply to development in general, i.e., what justifies characterising developmental processes as generally 'purposive', in this sense, when it is often not evident that they are conducing to a distinctly adaptive

morphological end state? To avoid this challenge, one could say that it is only *adaptive* developmental processes that are amenable to the proposed teleological or agential form of explanation, but, whenever organisms *do* develop adaptively, invoking concepts of agency and purposiveness still gives us an ineliminable insight into *why* they are responding to perturbations in the way they are.

We take it that this heuristic perspective is appropriate insofar as it goes. It does valuable epistemic work in highlighting the importance of developmental robustness in securing well adapted morphological outcomes. But we take issue with it on two fronts. First, we suggest that deploying the language of organismal agency to formulate this perspective is misleading. These concepts actively invite confusion, with reasonable readers primed to interpret discussion of the embryo's 'pursuit of its goals' as implying a literal, *directive* causal power, guiding development toward an adaptive end. This confusion is compounded by the frequent use of active verb constructions, often explicitly tagged with the modifier 'active', which frame the organism as the subject of development, actively "using", "marshalling", "altering", and "directing" its own genes, proteins, cells, tissues, and organs. This clearly does not correspond to the picture painted by the heuristic version of the agential stance, which, instead, depicts the organism's agency and purposiveness in its development as consisting in the (non-causal) relation of hypothetical necessity that holds whenever developmental processes are (robustly) adaptive. Supplementing this with distinctly causal language is needlessly misleading, if a literal interpretation is not what is intended.

Second, we argue that this approach does not offer any sort of distinctive or ineliminable explanation for *why development tends toward adaptive outcomes,* and thus why evolution is adaptively biased, beyond what is already entailed by the Darwinian logic underwriting existing evolutionary models and explanations. In order to do so, the embryo's agential 'pursuit of its goals' would somehow need to (literally) underlie or bring about the adaptive directionality of its development—and, consequently, of its lineage's evolution—in such a way that can be differentiated from the biasing effects of natural selection. Yet there seems little empirical support for this and, in any case, it is not what is claimed by the heuristic interpretation of the agential stance. On the contrary, as we show in the next section, the *adaptive* robustness—and, hence, the apparent 'goal-directedness'—of embryonic development, which is what makes development amenable to the relevant form of teleological explanation, can readily be explained as a selected trait. The tendency of developmental processes to robustly attain well-adapted morphological outcomes is therefore more parsimoniously explained as the passive consequence of the configuration

of the system, such that it has the statistical tendency to converge onto a viable phenotype—without the need to posit additional explanatory factors, such as the organism's active, real-time pursuit of its goal to persist.

6.5.1. Robustness and phenotypic plasticity are selected traits

In general, any complex physical dynamical system that persists far from thermodynamic equilibrium, across changeable environments, will tend to adopt configurations that confer robustness across diverse conditions (Chvykov et al., 2021; England, 2022). This includes the appearance of a sort of "memory" that lets the system rapidly shift to latent attractor states that were stable in previously encountered environments. This is consistent with the idea that complex systems that persist necessarily come to model aspects of their environment in their own physical structures (Conant & Ashby, 1970; Still et al., 2012). In effect, these tendencies reflect the outcome of a tautological selection process that selects simply for physical persistence. Where living systems are concerned, which do not just persist but also reproduce, natural selection will do the same job, by selecting embryonic configurations that can robustly generate viable individuals over a range of experienced environments (Wagner, 2013; Watson & Szathmáry, 2016; Szilágyi et al, 2020).

Indeed, there is a wealth of evidence supporting the idea that robustness is itself a selected trait of the developmental system (Alon, 2006; Wagner 2013). First, there is the well-known and general observation that phenotypes resulting from genetic mutation are not just different from the wild-type phenotype, but *more variable*. This was noted by Conrad Waddington already in 1957. He also observed that some genotypes are associated with greater levels of developmental robustness than others:

> "Another essential point about histogenesis (and morphogenesis) is that the *degree* of canalization is under genetic control. That is to say, individuals of some genotypes show a more powerful tendency to regulate to the normal canalised paths of development than do others." (p.20)

It is even possible to experimentally screen for mutant genotypes that specifically increase *variance* of a quantitative phenotype without altering the mean value (Ayroles et al., 2015; Mestek-Boukhibar & Barkoulas, 2015; Hallgrimsson et al., 2019). Different genotypes are thus linked to differing levels of developmental robustness, and, consequently, different amounts of inherent developmental variability (manifest as phenotypic variability across clones of genetically identical individuals reared in the same environments) (Vogt, 2015; Mitchell, 2018a).

Similarly, the capacity to produce alternate, adaptive phenes in response to varying environmental conditions or factors is a species-specific trait which is selected for and, again, appears to have a genetic basis (Pigliucci, 2005; Lea et al., 2018). Many closely related species to those with celebrated examples of phenotypic plasticity (e.g., *Daphnia*, locusts, and aphids) *do not show* such plasticity when exposed to the same environmental conditions or factors. This makes sense, in that any such capacity must rely on some specifically configured biochemical and cellular components and the network of regulatory interactions between them that actually mediate the developmental processes involved in producing the alternate phenotypes (Beldade et al., 2011). There is thus every reason to expect that observed instances of adaptive phenotypic plasticity reflect prior selection for that capacity (Pigliucci, 2005; Ghalambor et al., 2015). Hence, these capacities are more akin to a pre-configured control 'policy' *implemented at the organism-level*, than an active or novel real-time 'decision' made *by the organism*. The selected nature of these capacities therefore does not support the view that organisms can agentially produce genuinely novel, *adaptive* phenotypes that could be the substrate for a distinctive form of "phenotype-first" evolution.

Instead, these observations provide strong evidence that the robustness of an organism's development (i.e., its apparent 'goal-directedness'), whether attained through adaptive plasticity or through canalization, is the consequence of a selection process with a predominantly genetic basis. Natural selection favours not just molecular systems that can mediate specific developmental processes, but systems that do so robustly. This is evident in the gene regulatory network and signal transduction motifs employed in multicellular development, where the specific subset of possible motifs actually observed are notably robust, in engineering terms (Alon, 2006). And it is evident in the observed distributed robustness of networks at multiple cellular and physiological levels (Payne et al., 2014; Félix & Barkoulas, 2015; Nijhout et al., 2017).

We therefore do not need to appeal to the organism's purposiveness, or its active pursuit of a goal, in any sense, to explain the observation that its developmental processes resiliently tend toward adaptive outcomes. Doing so not only faces some conceptual difficulties with regard to non-adaptive and maladaptive developmental outcomes, but it also obscures the fact that existing selectionist models of adaptation already have the conceptual toolkit for more parsimonious and better empirically supported explanations of why developmental processes, such as canalization and plasticity, often tend toward adaptive outcomes (rather than other possible outcomes).

At best, then, the agential stance provides a new vocabulary for *describing* certain (holistic) processes that are necessary to realise or enact the adaptive tendencies of development and evolution. However, this does not translate into any sort of novel, radical, or improved understanding of *why* these processes exhibit this adaptive character (as opposed to being non-adaptive or maladaptive), as is often claimed. We therefore contend that, while a literal interpretation of the agential stance is untenable, a heuristic interpretation is simply not strong enough to support the implications of (ii) *agential adaptation* and (iii) *agential evolution*. This view appears to merely re-present the well-established passive 'goal' of development (i.e., a statistical tendency to produce a well-adapted individual organism) as an active 'goal' of the developing organism (i.e., a real-time pursuit of adaptation), without sufficient justification for doing so. If one's object of explanation is the *adaptive bias* observed in both development and evolution, then we suggest that standard selection-based approaches are more parsimonious and already appear entirely consistent with the purported insights of the agential stance on development.

6.6. Organism-Level Control

What about the implication outlined in (i) *Organism-Level Control*? If one's object of explanation is the *causes of morphological development*, then some of the vocabulary afforded by the agential stance may indeed carry some noteworthy, *prima facie* strengths. First, it explicitly foregrounds the essential role that emergent levels of organisation (including the whole-organism level) play in altering the causal landscape of the developing system and, thus, in securing well-adapted morphological outcomes (when they occur). Second, it strongly emphasises the robustness—or equifinality—of adaptive development, which is often missed from more reductionist, typically gene-focused, perspectives.

We therefore do not disagree that there is value in adopting an *organism-centric* perspective on development (see Walsh, 2018). Our concern is that, for the reasons given above, framing this organism-centric perspective in active, agential language often obscures the fact that this is only one perspective (or level) at which the multi-scale causal dynamics of development can be analysed. Both implicitly and explicitly, it gives the impression of an argument *against* other perspectives (e.g., gene-centrism or genotype-first adaptationism) and an argument for endorsing 'top-level causation' *over* 'multi-scale causation'.

By contrast, we would contend that the lesson from complexity science is that there is no 'privileged' level *or* timescale of causality (at least with regard to the processes of development)—no 'executive controller' at all (genetic, organismic, or otherwise). Observations of plastic self-organisation do not only imply evidence of a lack of an 'external director', they also speak to the absence of an 'internal director'. The whole point is that order emerges from the dynamics of a collective, as a 'happening', not a 'doing'.

As Susan Oyama (2000) describes in the Developmental Systems Theory perspective:

> "Form emerges in successive interactions. Far from being imposed by some agent, it is a function of the reactivity of matter at many hierarchical levels, and of the responsiveness of those interactions to each other." (p.22)

We therefore suggest that the language of agency, purpose, and goals is not well suited to the problem at hand. Instead, we should indeed recognise that development is indeed a holistic, whole-system, non-decomposable set of complex processes. Organisms exhibit reactive dynamics of self-organisation which manifest (sometimes) as adaptive developmental robustness or phenotypic plasticity. These are system-level properties that are crucial to understanding the causes of morphological development, but that does not mean they require the exercise of active, real-time control *by a subject*. Rather, any complex physical system that persists through time, including hurricanes and candle flames, will exhibit precisely these same kinds of robust dynamics to environmental variation and other perturbations (Meena et al., 2023). In living systems, these dynamics explicitly reflect the past effects of natural selection in the way the zygote (and developmental environment) is configured such that these tendencies obtain, often in an adaptively biased manner. We therefore submit that it is entirely appropriate, and thus more parsimonious, to continue to think of them as passive tendencies (things that tend to happen) rather than capacities that need to be actively exercised by a subject (i.e., the 'doings' of an agent).

Chapter 7

Discussion

7.1. Overview

The idea that we are agents, capable of choosing and controlling how we behave on the basis of our own intrinsic goals, beliefs and intentions, is foundational to the phenomenology of our everyday existence as human beings, to our basic self-conception, and to the moral, legal, and social norms that structure our societies. How we think about and conceptualise this notion of agency—both in terms of *where* in nature it applies and in what form it applies—has demonstrable and sometimes profound effects on many areas of our day-to-day lives, from the way in which we treat each other to how we quite literally perceive the world around us.

However, this concept of agency is also often claimed to be at odds with a truly naturalistic understanding of the world, that is, with the picture of the universe (and of biological systems in particular) given to us by the natural sciences. Skeptics argue that agency and related notions of choice, meaning, and freedom carry with them certain metaphysical commitments that are simply incompatible with what our best scientific theories tell us the world is like.

Consequently, many philosophers and scientists have concluded that our intuitive sense of agency—on which so much hangs—must in fact be an *illusion*. They claim that we are not *really* free to choose and control how we behave; our actions are not *really* 'up to us'. It just feels like they are. On this basis, some have even called for a radical overhaul of the many social institutions and personal practices that presuppose such agency (e.g., Pereboom, 2014; Caruso, 2021; Sapolsky, 2023).

This thesis has attempted to push back on these conclusions by arguing that the concept of agency *can* in fact be reconciled with scientific naturalism. To make this problem tractable, I focused in on three specific features (or types) of agency that are commonly claimed to be in tension with the natural sciences. These were: (i) agent-level causal sourcehood, (ii) semantic causation, and (iii) genuine choice (or 'the ability to do otherwise'). These features provided the thematic scaffold that I then used to structure the underlying direction and scope of this research project, and the arguments therein.

In close collaboration with my supervisor, Kevin Mitchell, I approached the challenge of naturalising these features of agency through a series of independent but interrelated

articles, each addressing a different topic related to agency from within the philosophy of action (**Chapter 2**), behavioural neuroscience (**Chapter 3**), philosophy of free will (**Chapters 4 and 5**), and evolutionary developmental biology (**Chapter 6**).

Although each of the previous chapters was written to serve as a standalone article, and although they were all situated within quite diverse literatures (both theoretically and terminologically), the core argumentative strategy of each chapter in this thesis was the same: once we attend carefully to the empirical findings of modern physics, biology, neuroscience, or psychology, the apparent *antinomy* (or, in one case, *parsimony*) between agency and naturalism dissolves.

The thesis's guiding methodological commitment was therefore one of *non-eliminative naturalism*. It pursued a realist account of agency that is empirically well-motivated and does not appeal to any metaphysically extravagant notions; yet remains faithful to and actually seeks to *explain* the relevant features of agency, rather than explaining them *away* as useful fictions or re-engineering the concept of agency so as to remove the features entirely.

The result is that each thesis chapter follows a very similar argumentative trajectory. First, I identify a particular claim or assumption that is generally taken to be empirically well-supported (within the relevant literature), and which appears to be in tension with one or more of the features of agency listed in (i)-(iii). Then, I draw on empirical evidence to develop a novel conceptual framework for thinking about the relevant biological phenomenon which, I argue, overturns the original assumption and, instead, provides an entirely naturalistic and non-mysterious description of the targeted feature(s) of agency. **Table 3** gives an overview of how this argumentative template was implemented within each chapter of this thesis.

Ch.	Field	Target Phenomenon	Targeted Assumption	Argument Summary
2	Philosophy of Action	Basic action in complex systems	Agent causation is naturalistically implausible.	I draw from thermodynamics and systems biology to develop an eight-criterion framework which, I argue, collectively describes an entirely naturalistic and non-mysterious form of agent causation.
3	Behavioural Neuroscience	Neural causes of behaviour	Necessary and sufficient neural mechanisms can provide complete causal explanations of behaviour	I integrate philosophy of causation with recent neuroscientific work to argue for a more expansive view of causation in the brain, which includes macroscopic, constraint-based, and even semantic causation.
4	Philosophy of Free Will	Evolution of nonlinear, macroscopic systems	The evolution of the universe and of our neural processes is deterministic.	I survey a range of empirical and conceptual arguments within modern physics to argue that determinism does not hold at any level of description in our universe.
5	Philosophy of Free Will	Deliberative decision-making in humans	Undetermined choices and actions are mere matters of chance or luck.	I integrate the conclusions from previous chapters, and draw on additional evidence from neuroscience, psychology and decision theory, to critique existing formulations of the Luck Objection and develop an alternative account of deliberative decision-making on which decisions are both undetermined *and* under the agent's control.
6	Evolutionary Developmental Biology	Embryonic development	Organisms actively and purposively control their own development.	I appeal to the wider evidence base in developmental and evolutionary biology to argue that claims about purposive developmental agency—when interpreted literally—are empirically unsupported.

Table 3. Overview of each thesis chapter's methodology and argument.

The rest of this concluding chapter performs two tasks. First, I synthesise and summarise the main conclusions and contributions of this thesis, focusing in particular on the naturalised account of causal sourcehood, semantic causation, and genuine choice developed throughout the thesis. Then, in the final section, I consider some avenues for future research that could build on this work.

7.2. Conclusions and Contributions

A Naturalised Account of Agent-Level Causation and Causal Sourcehood

One of the core claims of this thesis is that biological systems like us really are the legitimate causal source of our own actions because we exhibit a macroscopic form of causal power that is not entirely reducible to microphysical laws and interactions (*contra* vertical reductionism) and that genuinely inheres at the level of the whole system (*contra* horizontal reductionism).

To motivate this claim, I first set out a roadmap for how such agent-level causation could be realised, at least in principle, within a complex theoretical system (**Chapter 2**). I presented an eight-part framework which, I argued, collectively describes a naturalistically plausible system that overcomes the challenges of causal reductionism and exhibits an entirely non-mysterious form of emergent, agent-level causation. There were two crucial moves in developing this argument. The first was the recognition that systems whose causal dynamics are *holistically integrated* are never truly amenable to a machine-like analysis that decomposes the system into its component parts and localises causal power to just *some* of those parts. The second was the argument that systems containing some *low-level indeterminacy* and a functional organisation in which the causal dynamics of the system are sensitive to *multiply realisable* macro-states (rather than any specific microphysical realisation of that state) genuinely exhibit a mode of irreducible, macroscopic causation.

Throughout the thesis, I then argued that these theoretical criteria for agent-level causation are satisfied, in practice, by many biological systems. In **Chapter 2**, for example, I showed how even in the case of one of the simplest known biological behaviours—bacterial chemotaxis—the organism is too *holistically integrated* to truly localise and isolate 'the cause' of that behaviour to a single, well-understood biochemical pathway *within* the system. Similarly, in **Chapter 6**, we saw how the functional organisation of embryos is also too deeply integrated to isolate the cause of (almost) any phenotypic outcome to the action of individual genes.

In **Chapter 4**, I argued extensively for the view that *low-level indeterminacy* really is a pervasive feature of all biological systems. And, in **Chapter 3**, I showed how neural systems are configured to be causally sensitive to *multiply realisable*, higher-order patterns (or equivalence classes) of neural activity, while being largely *insensitive* to the precise microphysical details of specific neuronal firings or ion flows.

Taken together, these arguments collectively support the view that natural systems, in theory, and certain biological systems, in practice, can genuinely possess the sort of emergent, macroscopic, agent-level causal power that many have traditionally held to be incompatible with a scientific worldview. In turn, this sort of naturalised account of agent-level causation provides the empirical resources required to defend the view that biological systems like us really *are* the true causal sources of our action, thereby deflating at least one of the standard naturalistic critiques of human agency.

This account of agent-level causation also comes with some wider payoffs and implications that are worth mentioning here, even if they were not the primary focus of this thesis. First, this account offers some novel conceptual and empirical resources that may be of use to agent-causal philosophers of action (e.g., Hornsby, 2004; Steward, 2012) and free will (e.g., O'Connor, 2000; 2009) in pushing back against the event-causal approaches that currently dominate these two fields.

Second, the eight-part framework presented here is intended to be both empirically grounded *and* operationalisable: thermodynamic autonomy, causal insulation, holistic integration, sensitivity to macrostates rather than microstates, and historicity are all potentially *measurable* properties of a system. This eight-part framework could therefore prove useful for testing and measuring the extent to which different systems (biological or artificial) exhibit macroscopic, agent-level causation, as a tractable means of then informing the emerging ethical debates and policy decisions surrounding AI responsibility (see **Chapter 1**). I return to this issue in **Section 7.3**.

Third and finally, an important feature of this eight-part framework is that it depicts agency as a multi-dimensional and graded phenomenon. Conceptually, this is very different from the binary, absolutist terms in which the concepts of agency and free will are typically discussed within philosophical circles. Instead, it suggests the possibility of systems having 'more or less' agency than one another—or, more plausibly, different 'agency profiles' from one another, due to inherent trade-offs between the different dimensions in this framework. If this is right, then it opens up interesting new ways of framing these debates and provides a potentially rich and fruitful alternative line of inquiry for philosophers of agency to pursue.

A Naturalised Account of Semantic Causation

A second key claim of this thesis is that certain biological systems really can act *for reasons*, in an ontologically robust sense. I aimed to show that it is entirely consistent with scientific naturalism to claim that the motivational states of a biological system—its purposes, beliefs, goals, values, and intentions—genuinely "make a difference" to how the system behaves, in virtue of their subjective meaning and semantic content (Froese & Taguchi, 2019).

To do this, I built on the insight that at least some biological systems are configured to be causally sensitive to *multiply realisable*, higher-order patterns (rather than their precise microphysical instantiations) in order to argue that, in many of these systems, the nature of these patterns (that is, the composition of their equivalence class) comes to embody and instantiate what is subjectively meaningful *to the organism*. If this is right, then—I argued—we would seem to have a firm naturalistic basis for something like the sort of semantic causation described above.

I motivated this claim on two fronts. First, in **Chapter 2**, I showed how natural selection initially, but then experience, learning and metacognition in more complex organisms, demonstrably tunes the configuration of these higher-order patterns to quite literally embody what is functionally useful or *meaningful* to the organism for achieving its goals (e.g., persistence). The key move here is recognising that living systems are fundamentally *processes* (Nicholson & Dupré, 2018; Meincke, 2018), with extension through time, and with a deep sensitivity to their own *historicity*—they 'carry their history on their backs' in a way that is often grounded in their own personal experience and interactions with the world, filtered through the lens of their purposes and goals, and physically structured into the causal architecture of the system itself.

Then, in **Chapter 3**, I furthered this argument by offering an empirically grounded account of how, in practice, meaning and subjectivity get built into the causal architecture of the brain. First, following Tse (2013), I argued for a constraint-based model of neural dynamics in which the macroscale organisation of the system constrains the inherently noisy microscale goings-on by configuring neurons and neural populations with coarse-grained, functional *criteria* to which they are causally sensitive. (This is how the brain implements its multiply realisable, higher-order patterns and becomes causally sensitive to them). Second, and crucially for *semantic* causation, I then noted that these organisational constraints—which are what ultimately *configure* the higher-order patterns—appear to be actively shaped by what Fred Dretske (1988) has called *structuring causes*: distal events, such as the subject's past experiences and historical interactions with

the world, which *structure* the system in a particular way, namely those ways that favour survival. The result is that the organism's personal historicity (via structuring causes that shape the system's organisational constraints in targeted ways) actively tunes neurons and neural populations to be causally sensitive to low-level neural dynamics which represent differences in the world that actually *matter* to the organism, and to be *insensitive* to those that do not. The outcome is a model of neural dynamics that gives causal priority to what neural states and neural activity *mean* to the organism, and *not* to the precise physical properties or instantiation of those states. On this view, neural activity only has causal power within the system *in virtue of what it means*—that is, in virtue of its semantic content.

There is of course still a lot of conceptual and philosophical work left to do to convincingly show how this account of semantic causation can connect to (and maybe even inform) the more traditional philosophical discussions surrounding mental causation and the mind-body problem. However, by developing an empirically grounded model of the neural causes of behaviour which centres the personal historicity of the organism, foregrounds what neural states *mean* to the organism, and backgrounds the exact physical realisation of those states, this thesis has contributed a promising new avenue through which these age-old problems might be addressed. It also provides an—albeit unfinished—rebuttal to Libet-style naturalistic critiques of agency that cast our mental states (our beliefs, desires and intentions) as mere epiphenomena.

Again, there are also some wider payoffs to this account of semantic causation that are worth briefly mentioning. First, in making the naturalistic case for semantic causation, **Chapter 3** introduced a suite of more expansive causal concepts for thinking about causation within the brain. If the arguments in this chapter are right, then this new causal schema seems to provide exactly the sort of multiscale, diachronic conceptual toolkit that will be needed to support the growing field of systems neuroscience, as it looks to take advantage of recent advances in brain imaging technology (see **Chapter 1**). This work could therefore prove beneficial to neuroscientists working in this field.

Second, in developing a more 'structural', 'internalist' (or even 'mechanistic') account of semantic causation, this thesis goes beyond the dominant 'ecological' model of purposive agency developed by Walsh (2015) and colleagues, by providing a potentially productive means through which to distinguish truly goal-directed, purposive systems from robustly end-oriented ones (e.g., hurricanes and embryos) (see **Chapter 6**). This is fast becoming an ethically important skillset for society to possess, given that current AI systems (e.g., LLMs) are effectively being trained to mimic human purposiveness and thus make it as

difficult as possible for us to intuit this distinction between genuine and merely apparent purposiveness.

A Critique of Classical Determinism and 'Determinism-Plus-Randomness'

The third overarching aim of this thesis was to develop and defend a naturalised account of genuine choice: i.e., the idea that it is (i) open to an agent to take more than one possible course of action, and (ii) under the agent's control which of these possible actions it takes. As outlined in **Chapter 1**, there is a long-standing, two-pronged challenge to the scientific credibility of genuine choice in the free will literature, which combines a challenge from determinism and a challenge from "luck" (or indeterminism).

Written in collaboration with physicist George Ellis, **Chapter 4** took the first steps towards addressing this two-pronged challenge by arguing, first, that determinism does not accurately describe our universe or our decision-making processes (and thus should not be considered a naturalistic threat to genuine choice); and, second, that the model of indeterminism traditionally assumed in the free will literature, and that typically underwrites this threat from luck, is not empirically well-supported.

For the first of these arguments, I drew on a range of empirical and conceptual work from across contemporary physics to make the case that determinism does not hold at any scale of physical reality. At the quantum level, I showed how the Heisenberg Uncertainty Principle, in conjunction with the phenomenon of "zero point energy", and the apparently random "collapse" of quantum systems under measurement, strongly undermine the plausibility of determinism at this scale of reality (see also **Chapter 2**). At the classical level, I showed how there is no empirical basis for thinking—as many do—that classical (or just macroscopic) scales of reality are somehow *insulated* from this indeterminacy at quantum scales, in a way that would support the thesis of classical determinism. On the contrary, I showed that there are several compelling reasons to think that classical determinism cannot hold true for non-linear, chaotic, macroscopic systems like us, *in principle*. I therefore concluded that determinism does not seem to pose the threat to free will (and to genuine choice) that much of the free will discourse would suggest it does.

Chapter 4's second key contribution was the claim that this naturalistic critique of determinism actually applies equally to the standard conception of *in*determinism within the free will literature. The reason, I argued, was that, while libertarian philosophers (and their critics) openly embrace *quantum* indeterminacy in their work, they still generally hold onto the tacit assumption that determinism remains "Nature's default mode" at

classical scales (Earman, 2008, p.817). Consequently, these philosophers are often forced to adopt a model of indeterminism in which metaphysical openness at the (macro-)level of decisions and action ultimately must derive from random quantum events 'percolating up' to disrupt *otherwise deterministic* goings-on in the brain and body—a model of indeterminism we termed 'determinism-plus-randomness'. **Chapter 4**'s second core argument therefore was that: if 'determinism-plus-randomness' presupposes (and, in fact, derives from) the thesis of classical determinism, and we have good empirical reason to reject classical determinism as an accurate description of nonlinear, macroscopic systems like us, then we have good reason to reject determinism-plus-randomness as the appropriate model of indeterminism for debates surrounding genuine choice and free will in humans.

A Novel Model of Indeterminism: Pervasive Indefiniteness

In place of determinism-plus-randomness, this thesis also proposed a novel—more empirically plausible—model of indeterminism from which to assess the naturalistic credibility of genuine choice. In **Chapter 4** (and, to some extent, in **Chapter 5**), I re-used the evidence from physics, which I had initially used to critique classical determinism and determinism-plus-randomness, to help motivate and develop an alternative, *positive* conception of indeterminism, which we labelled 'pervasive indefiniteness'.

Under this view, indeterminism is conceptualised primarily as a *negative*, as a basic and general *under*-determination of the future, rather than as a *positive* addition of truly random events to an *otherwise deterministic* universe.

As argued in **Chapter 4** and put into practice in **Chapter 5**, if this really is the right way to think about the fundamental metaphysics of living systems, then it would seem to encourage a substantial reframing of many of the core questions and topics that currently dominate the free will literature. To give one example, if it really is naturalistically unlikely that determinism accurately describes nonlinear, macroscopic systems like us, then it would seem inappropriate to persist with a theoretical terrain in which "the basic divide among philosophers is between compatibilism and incompatibilism" (Palmer, 2014, p.4), where the main disagreement is over the implications of determinism for free will. A more empirically aligned literature would, instead, primarily distinguish between theorists who see free will as compatible with indeterminism (e.g., libertarians and two-way compatibilists) and those who do not (e.g., hard incompatibilists and one-way compatibilists).

Further, if pervasive indefiniteness really is the most empirically plausible model of indeterminism, then it seems to provide a new conceptual foundation with which to revisit the second (and only remaining) horn of the argument against genuine choice: the challenge from "luck". This was the subject of **Chapter 5.**

A Naturalised Account of Genuine Choice that Overcomes the Luck Objection

Having deflated determinism and proposed a novel model of indeterminacy, I then used this new metaphysical framework to defend the thesis's final core feature of agency—genuine choice—from the threat from luck (**Chapter 5**). To do this, I took a dual approach.

First, I argued that most formulations of the Luck Objection either presuppose a model of deliberative decision making which we have good scientific reason to think does not apply to biological systems like us *or* do not successfully establish that indeterminacy strips agents of the sort of control required to genuinely choose what one does. With regard to the former, I argued that three of the most popular versions of the Luck Objection—the Reductive Luck Objection, the 'Objective Probabilities' Luck Objection, and the 'Disappearing Agent' Luck Objection—rely on features of indeterministic decision making (determinism-plus-randomness, prior objective probabilities, and event causation, respectively) which are empirically implausible as features of *human* deliberative processes. With regard to the latter, I argued that two of the most prominent *explanatory* versions of the Objection—the Contrastive Luck Objection and the Problem of Present Luck—tell us something important about the limits of agential control (namely, that it is not *absolute*), but do not show that it is not still sufficiently under the agent's control *which* of the available courses of action they end up taking. I therefore concluded that we have good *negative* reasons to reject the challenge from luck for genuine choice, insofar as none of the most prominent formulations of the challenge seem to be successful in their task.

The second aspect of my dual approach to naturalising genuine choice was then to pursue the more positive line of argument. In making the empirical case *against* these various versions of the Luck Objection, I simultaneously built up and defended a more empirically plausible *alternative* model of indeterministic decision making in humans, one which integrated several of the key ideas and arguments from my previous chapters (agent-level causation (**Chapter 2**), constraint-based control (**Chapter 3**), and pervasive indefiniteness (**Chapter 4**)). On this view, human agents exercise a form of macroscopic, agent-level control over their decision-making processes by constraining the inherently noisy goings-on within them, in a way that is guided by their own motivations and real-time reasoning

processes, and which gradually leads to the system-as-a-whole settling into a new global dynamical state—a dynamical state that just *is* the agent coming to favour and, ultimately, choosing one of the available courses of action. Crucially for the concept of genuine choice, none of this is pre-determined or even pre-statable, at either the physical or the psychological level. Instead, on this model, agents simply have to *figure out*, in real-time, as best they can, which of the available options to favour and then *do work* to bring forth a future that satisfactorily aligns with that preference, using noisy components and under uncertain conditions. The result is an empirically grounded model of deliberative decision making in which an agent's decisions are neither pre-determined nor a random happening. Instead, it is open to the agent, prior to deliberation, to choose more than one action and then, during deliberation, the agent exercises her reasons-guided, agent-level control to actively constrain the possibility space of what *could* happen in order to (non-deterministically) bring about one of those particular actions; namely, the one she comes to favour during deliberation.

If this is an appropriate description of our decision-making process, then these arguments would seem to entail an entirely naturalistic and non-mysterious account of genuine choice, thereby defending the empirical credibility of this thesis's third and final core feature of agency.

A Critique of Developmental Agency and its Implications for Evolutionary Theory

The final contribution of this thesis was the argument that (i) developing organisms are *not* purposive agents, actively directing and controlling their own embryonic development, in a literal sense, and thus (ii) evolutionary theory is *not* in need of the sort of radical overhaul that many in theoretical biology have recently been calling for (e.g., Walsh, 2015; Sultan et al., 2022; Nadolski & Moczek, 2023) (**Chapter 6)**.

I defended this claim in two stages. First, I argued that—contrary to what is commonly suggested in this literature—the empirical evidence in developmental biology does *not* support a literal ascription of purposive developmental agency to developing organisms. I show how the arguments given to support this view appear to rely on a biased sampling of the developmental evidence base and that, once a more even-handed survey of the evidence is taken into consideration, the intuitive basis for these arguments disappears.

The second stage of my argument then showed how, if these claims of purposive developmental agency are *not* intended literally (as some have claimed (e.g., Fulda, 2023)),

then they do not support the radical implications for evolutionary theory often argued to follow from them. The reason, I suggest, is that, when intended heuristically (i.e., non-literally), describing a system as exhibiting purposive developmental agency is just another way of saying that the system is robust to perturbation; it can continue progressing toward some end state, even if it is knocked off track. Yet, I show that current evolutionary theory already accounts for (and, moreover, it has the resources to *explain*) this robustness of developmental systems. Thus, I concluded, a non-literal ascription of purposive developmental agency does no distinctive or ineliminable explanatory work that cannot already be done within current evolutionary theory. Hence, the claim that developing organisms are purposive agents cannot and should not be used to motivate radical re-formulations of evolutionary theory.

Within the context of the rest of this thesis, the arguments presented in **Chapter 6** also serve an important rhetorical function. They illustrate that the methodology of *non-eliminative naturalism* that has guided my investigation of agency throughout this thesis 'cuts both ways', so to speak. It is not just a tool with which to hand-wavingly identify 'naturalised' agency everywhere in nature. It encourages an empirically constrained and tempered approach to naturalising agency—as I think **Chapter 6** illustrates.

Summary

In summary, this thesis defends a naturalised picture of agency on three main fronts. First, agent-level causal sourcehood is secured via an eight-criterion framework, under which complex agents are argued to exert a macroscopic form of control that is emergent and genuinely inheres at the level of the whole system, thereby resisting concerns about causal reductionism. Second, semantic causation is naturalised (in neural systems, at least) by showing how neural dynamics come to be causally sensitive to—and, ultimately, driven by—what is subjectively meaningful to these systems rather than by their physical details *per se*, thereby resisting the threat of eliminative materialism. Third, genuine choice is shown to be compatible with science by (a) taking seriously the empirical evidence against determinism, (b) replacing the prevailing "determinism-plus-randomness" model of *in*determinism with a more empirically plausible "pervasive indefiniteness" model, and (c) recognising deliberative decision-making as a temporally extended, noisy, but still agentially controlled process that is neither pre-determined nor random. Together, these arguments provide a framework for thinking about complex biological systems and their behaviour that reconciles agency with scientific naturalism without eliminating or redefining any of the concept's most intuitive core features. The thesis's final chapter then

used this framework to evaluate some recent influential attempts to ascribe agency to developing organisms, arguing that this is not successful.

In the next section, I consider some potential directions for future research that could build on the conclusions and contributions of this thesis.

7.3. Future Directions

Ultimate responsibility, belief formation, and semantic causation

In this thesis, I aimed to contribute to the project of naturalising agency by focusing on three of the concept's core features: causal sourcehood, semantic causation, and genuine choice. However, these are certainly not the only features of our intuitive concept of agency that have been the subject of naturalistic scrutiny. The project of naturalising agency could therefore benefit from future research that applies the same approach of non-eliminative naturalism employed in the previous chapters to other naturalistic concerns about agency.

One major example would be the notion of *ultimate responsibility*. Briefly, ultimate responsibility refers to the idea that in order for our actions to be truly considered 'up to us', we not only need to choose and control how we behave on the basis of intrinsic reasons and values, we also need to somehow be *responsible* for having those reasons and values in the first place. That is, we need to be the source or creator of our own psychological character if we are to be the true causal source of the actions that, in some sense, issue from that character.

Many contend that this notion of ultimate responsibility is foundational to the way in which we currently attribute (moral) agency to one another. Aristotle, for example, noted that "*if a man is responsible for wicked acts that flow from his character, he must at some time in the past have been responsible for forming the wicked character from which these acts flow*" (1985, p.63).

Skeptics about agency, however, commonly argue that ultimate responsibility is not a naturalistically plausible feature of living systems. They claim—as Schopenhauer (1839/1999) did in his infamous (albeit dated) line—that, under a scientific worldview, 'man can do what he wants but not want what he wants' (see also Nietzsche, 1886, section 21; Strawson, 1994). The apparent reason for this is that:

> "You did not pick your parents or the time and place of your birth. You didn't choose your gender or most of your life experiences. You had no control whatsoever over your genome or the development of your brain. And now your

> brain is making choices on the basis of preferences and beliefs that have been hammered into it over a lifetime—by your genes, your physical development since the moment you were conceived, and the interactions you have had with other people, events, and ideas. Where is the freedom in this? Yes, you are free to do what you want even now. But where did your desires come from?" (Harris, 2012, p.41)

Future research could therefore use the same non-eliminative, naturalistic approach employed throughout this thesis to extend the project of naturalising agency by asking: can we reconcile the notion of ultimate responsibility with scientific naturalism?

One promising avenue for exploring this question would be to draw on the neuroscientific and psychological literature on belief formation and goal acquisition as a means of investigating Harris' core assumption here that our preferences and beliefs are simply 'hammered into us' by external events in a way that affords us no control or authorship over them (i.e., no ultimate responsibility). Indeed, there is certainly some evidence to suggest that the way we acquire beliefs is better understood as being filtered through and often actively mediated by our current psychological state (Bower, 1981; Kunda, 1990; Sharot et al., 2011; Seth, 2013; Seitz & Angel, 2020; Albarracin & Pitliya, 2022). If this is right, then it seems to put pressure on the idea that our preferences and beliefs are simply being 'hammered into us'. This would therefore be an interesting avenue to explore for future research looking to naturalise ultimate responsibility.

In doing so, this research may also help to address one of the key outstanding questions from the account of semantic causation I have given in this thesis. That is because it would seem to provide a fruitful opportunity to explore how exactly it is that the meaning on which a neural system runs—i.e., the composition of the higher-order *patterns* to which the system is causally sensitive—can come to be shaped by something *other* than the system's pursuit of its own persistence. The potential for meaning to be tethered to more everyday goals or purposes was touched on in **Chapter 3** (in the distinction between pragmatic and semantic causation) but a detailed biomechanistic account of *how* this could happen was not given. Future research on ultimate responsibility could therefore also prove beneficial for fleshing out this specific limitation of the account of semantic causation developed in this thesis.

Formalising agency as a measurable, graded, multi-dimensional concept for use in AI ethics

The previous chapters sought to develop a naturalistic metaphysics for thinking about living systems and their behaviour which supports an entirely non-mysterious and

empirically plausible account of biological agency (at least along three key dimensions). Along the way, I have repeatedly flagged the fact that, on this account, agency appears to come out as a graded, multi-dimensional property of biological systems, not a binary or categorical one. I also noted that, at least in theory, this account seems to provide quantifiable hallmarks that support the possibility of formalising agency as a tractably measurable property of systems (at least to some degree).

Throughout the thesis, I gestured toward several of these potentially operational measures of agency, but did not explicitly undertake the task of translating the proposed conceptual-metaphysical framework into a directly empirically testable one. Future research could therefore build on the arguments presented in this thesis by connecting them to existing (and maybe even developing new) operational measures. This would provide the crucial final step in establishing agency as a functional and useful *scientific* concept that could be applied and tested within experimental contexts.

A promising direction for this work to take would be to identify candidate quantitative metrics for each of the eight-criterion framework that I developed in **Chapter 2** (and which re-appear throughout the rest of the thesis). To give just one example, there is a growing selection of information-theoretic measures quantifying the extent to which coarse-grained macroscale descriptions of a system provide more information about the system's causal dynamics than even the most detailed microscale description of the same system (Hoel et al., 2013; Hoel, 2017, 2018, 2025; Albantakis et al., 2019; Rosas et al., 2020; Barnett & Seth, 2023; Rosas et al., 2024). These measures therefore seem to be capturing the extent to which the system is causally sensitive to multiply realisable *macro-states* (or higher-order *patterns*) rather than merely the specific microphysical realisation of that macro-state/pattern at a given time (Hoel, 2017; 2025). These measures therefore seem very well-placed to capture and operationalise the notion of macroscopic, informational causation that serves as the conceptual basis for the account of semantic causation defended in this thesis.

Developing a *measurable* account of biological agency in this way could prove to be of profound importance for the ethics of AI. It would provide a means by which to test for artificial agency within AI systems and thereby grapple more productively with a host of emerging ethical questions surrounding the rise of AI, including questions about the moral patiency and even moral responsibility of these AI "agents" (see **Chapter 1**).

7.4. Concluding Remarks

In conclusion, if the arguments in this thesis are sound, then we need not choose between a scientifically respectable worldview and a conception of ourselves as causal agents, acting for reasons. The concept of agency can live on—not as an illusion or a mystery, but as an empirically plausible description of the complex causal architecture of evolved biological systems. On this view, it is perfectly compatible with naturalism to say that agents *themselves* are causal sources in the world, that their subjectivity and meaning is causally efficacious, and that their choices are genuinely open and yet under their control. This reconciliation is not purely existential, either: it safeguards our self-conception and societal institutions from skeptical, revisionary challenges, it lays the foundations for an empirically tractable concept of agency that can be put to use in the life sciences and AI research, and it provides the basis for a continued research program in naturalising agency that encourages collaboration from across philosophy, biology, and psychology.

Bibliography

Abbott, L. F., & Regehr, W. G. (2004). Synaptic computation. *Nature*, *431*(7010), 796-803. https://doi.org/10.1038/nature03010

Abellán, C., Acín, A., Alarcón, A., Alibart, O., Andersen, C. K., Andreoli, F., … & BIG Bell Test Collaboration. (2018). Challenging local realism with human choices. *Nature*, *557*(7704), 212-216. https://doi.org/10.1038/s41586-018-0085-3

Acín, A., & Masanes, L. (2016). Certified randomness in quantum physics. *Nature*, *540*(7632), 213-219. https://doi.org/10.1038/nature20119

Adler, S. L. (2003). Why decoherence has not solved the measurement problem: a response to P.W. Anderson. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *34*(1), 135-142. https://doi.org/10.1016/S1355-2198(02)00086-2

Aguilar, J. H., & Buckareff, A. A. (2010). *Causing Human Actions: New Perspectives on the Causal Theory of Action*. The MIT Press.

Albantakis, L. (2021). Quantifying the Autonomy of Structurally Diverse Automata: A Comparison of Candidate Measures. *Entropy*, *23*(11), 1415.

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. *Entropy*, *21*(5), 459.

Albarracin, M., & Pitliya, R. J. (2022). The nature of beliefs and believing [Opinion]. *Frontiers in Psychology*, *Volume 13 - 2022*. https://doi.org/10.3389/fpsyg.2022.981925

Alberch, P. (1989). The logic of monsters: evidence for internal constraint in development and evolution. *Geobios*, *22*, 21-57. https://doi.org/10.1016/S0016-6995(89)80006-3

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits* (First Edition). Chapman & Hall/CRC. https://doi.org/10.1201/9781420011432

Alvarez, M. (2013). Agency and Two-Way Powers. *Proceedings of the Aristotelian Society*, *113*(1.1), 101-121.

Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: exploring agency within AI systems. *AI & Society*, *38*(4), 1321-1331. https://doi.org/10.1007/s00146-021-01326-6

Anscombe, G. E. M. (1971). *Causality and determination: an inaugural lecture*. Cambridge University Press.

Aristotle. (1985). *The Nicomachean Ethics* (T. Irwin, Trans.). Hackett.

Atri, D., & Melott, A. L. (2014). Cosmic rays and terrestrial life: A brief review. *Astroparticle Physics*, *53*, 186-190. https://doi.org/10.1016/j.astropartphys.2013.03.001

Ayroles, J. F., Buchanan, S. M., O'Leary, C., Skutt-Kakaria, K., Grenier, J. K., Clark, A. G., Hartl, D. L., & de Bivort, B. L. (2015). Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*, *112*(21), 6706-6711. https://doi.org/10.1073/pnas.1503830112

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, *22*(2), 170-188.

Baedke, J. (2021). What's Wrong with Evolutionary Causation? *Acta Biotheoretica*, *69*(1), 79-89. https://doi.org/10.1007/s10441-020-09381-0

Băetu, T. (2015). When Is a Mechanistic Explanation Satisfactory? Reductionism and Antireductionism in the Context of Mechanistic Explanations. In I. D. Toader, G. Sandu, & I. Pârvu (Eds.), *Romanian Studies in Philosophy of Science*. Springer Verlag.

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. In *Advances in child development and behavior, Vol. 30.* (pp. 153-187). Academic Press.

Balaguer, M. (2010). *Free Will as an Open Scientific Problem*. The MIT Press.

Ball, P. (2006). *Critical mass: How one thing leads to another*. Farrar, Straus and Giroux.

Ball, P. (2023a). Organisms as agents of evolution. *John Templeton Foundation.*

Ball, P. (2023b). *How Life Works: A User's Guide to the New Biology*. Picador.

Bandak, D., Mailybaev, A. A., Eyink, G. L., & Goldenfeld, N. (2024). Spontaneous Stochasticity Amplifies Even Thermal Noise to the Largest Scales of Turbulence in a Few Eddy Turnover Times. *Physical review letters*, *132*(10), 104002. https://doi.org/10.1103/PhysRevLett.132.104002

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191-215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122-147. https://doi.org/10.1037/0003-066X.37.2.122

Bandura, A. (2006). Toward a Psychology of Human Agency. *Perspectives on Psychological Science*, *1*(2), 164-180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359-371. https://doi.org/10.1038/s41583-021-00448-6

Barack, D. L., Miller, E. K., Moore, C. I., Packer, A. M., Pessoa, L., Ross, L. N., & Rust, N. C. (2022). A call for more clarity around causality in neuroscience. *Trends in Neurosciences*, *45*(9), 654-655. https://doi.org/10.1016/j.tins.2022.06.003

Barandes, J. A. (2023). The Stochastic-Quantum Theorem. *arXiv preprint*. https://doi.org/10.48550/arXiv.2309.03085

Barandes, J. A. (2025). The Stochastic-Quantum Correspondence. *Philosophy of Physics*, *3*(1), 8.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, *17*(5), 367-386. https://doi.org/10.1177/1059712309343819

Barnett, L., & Seth, A. K. (2023). Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems. *Physical Review E*, *108*(1), 014304. https://doi.org/10.1103/PhysRevE.108.014304

Barsalou, L. W. (2008). Grounded Cognition. *Annual review of psychology*, *59*(Volume 59, 2008), 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bartsch, J. C., Schott, B. H., & Behr, J. (2023). Hippocampal Dysfunction in Schizophrenia and Aberrant Hippocampal Synaptic Plasticity in Rodent Model Psychosis: a Selective Review. *Pharmacopsychiatry*, *56*(02), 57-63. https://doi.org/10.1055/a-0960-9846

Barwich, A.-S. (2021). Imaging the living brain: An argument for ruthless reductionism from olfactory neurobiology. *Journal of Theoretical Biology*, *512*, 110560. https://doi.org/10.1016/j.jtbi.2020.110560

Baumann, R. (2005). Soft errors in advanced computer systems. *IEEE Design & Test of Computers*, *22*(3), 258-266. https://doi.org/10.1109/MDT.2005.69

Bekenstein, J. D. (2004). Relativistic gravitation theory for the modified Newtonian dynamics paradigm. *Physical Review D*, *70*(8), 083509. https://doi.org/10.1103/PhysRevD.70.083509

Beldade, P., Mateus, A. R. A., & Keller, R. A. (2011). Evolution and molecular mechanisms of adaptive developmental plasticity. *Molecular ecology*, *20*(7), 1347-1363. https://doi.org/10.1111/j.1365-294X.2011.05016.x

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, *1*(3), 195-200. https://doi.org/10.1103/PhysicsPhysiqueFizika.1.195

Ben-Yami, H. (2020). The Structure of Space and Time, and the Indeterminacy of Classical Physics. *arXiv preprint*. https://doi.org/10.48550/arXiv.2005.05121

Bickle, J. (2015). Marr and Reductionism. *Topics in Cognitive Science*, *7*(2), 299-311. https://doi.org/10.1111/tops.12134

Bishop, J. (1989). *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.

Blouw, P., Solodkin, E., Thagard, P., & Eliasmith, C. (2016). Concepts as Semantic Pointers: A Framework and Computational Model. *Cognitive Science*, *40*(5), 1128-1162. https://doi.org/10.1111/cogs.12265

Boden, M. A. (2008). Autonomy: what is it? Introduction. *Biosystems*, *91*(2), 305-308. https://doi.org/10.1016/j.biosystems.2007.07.003

Boekholt, T. C. N., Portegies Zwart, S. F., & Valtonen, M. (2020). Gargantuan chaotic gravitational three-body systems and their irreversibility to the Planck length. *Monthly Notices of the Royal Astronomical Society*, *493*(3), 3932-3937. https://doi.org/10.1093/mnras/staa452

Born, M. (1969). Is Classical Mechanics in Fact Deterministic? In M. Born (Ed.), *Physics in My Generation* (pp. 78-83). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-25189-8_7

Bossaerts, P., Yadav, N., & Murawski, C. (2019). Uncertainty and computational complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1766), 20180138. https://doi.org/10.1098/rstb.2018.0138

Boué, G., Laskar, J., & Farago, F. (2012). A simple model of the chaotic eccentricity of Mercury. *A&A*, *548*, A43. https://doi.org/10.1051/0004-6361/201219991

Bower, G. H. (1981). Mood and memory. *American Psychologist*, *36*(2), 129-148. https://doi.org/10.1037/0003-066X.36.2.129

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263-1268. https://doi.org/10.1038/nn1525

Brembs, B. (2021). The brain as a dynamically active organ. *Biochemical and Biophysical Research Communications*, *564*, 55-69. https://doi.org/10.1016/j.bbrc.2020.12.011

Buchak, L. (2013). Free Acts and Chance: Why the Rollback Argument Fails. *The Philosophical Quarterly*. https://doi.org/10.1111/j.1467-9213.2012.00094.x

Budaev, S., Jørgensen, C., Mangel, M., Eliassen, S., & Giske, J. (2019). Decision-Making From the Animal Perspective: Bridging Ecology and Subjective Cognition [Hypothesis and Theory]. *Frontiers in Ecology and Evolution*, *Volume 7 - 2019*. https://doi.org/10.3389/fevo.2019.00164

Buzsáki, G. (2006). *Rhythms of the Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, *68*(3), 362-385. https://doi.org/10.1016/j.neuron.2010.09.023

Buzsáki, G. (2019). *The Brain from Inside Out*. Oxford University Press. https://doi.org/10.1093/oso/9780190905385.001.0001

Buzsáki, G. (2020). The Brain-Cognitive Behavior Problem: A Retrospective. *eNeuro*, *7*(4). https://doi.org/10.1523/eneuro.0069-20.2020

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

Carroll, S. M. (2024). *Quanta and Fields: The Biggest Ideas in the Universe*. Penguin Publishing Group.

Carroll, S. M. (2021). Consciousness and the Laws of Physics. *Journal of Consciousness Studies*, *28*(9-10), 16-31.

Caruso, G. D. (2012). *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will*. Lexington Books.

Caruso, G. D. (2014). Précis of Derk Pereboom's Free Will, Agency, and Meaning in Life. *Science, Religion and Culture*, *1*(3), 178-201.

Caruso, G. D. (2021). *Rejecting retributivism: Free will, punishment, and criminal justice*. Cambridge University Press.

Caruso, G. D., & Dennett, D. C. (2021). *Just Deserts: Debating Free Will*. Polity.

Caruso, G. D., List, C., & Clark, C. J. (2020). Free Will: Real or Illusion - A Debate. *The Philosopher*, *108*(1).

Cashmore, A. R. (2010). The Lucretian swerve: The biological basis of human behavior and the criminal justice system. *Proceedings of the National Academy of Sciences*, *107*(10), 4499-4504. https://doi.org/10.1073/pnas.0915161107

Castaneda, A. N., Huda, A., Whitaker, I. B. M., Reilly, J. E., Shelby, G. S., Bai, H., & Ni, L. (2024). Functional labeling of individualized postsynaptic neurons using optogenetics and trans-Tango in Drosophila (FLIPSOT). *PLOS Genetics*, *20*(3), e1011190. https://doi.org/10.1371/journal.pgen.1011190

Chalmers, D. J. (2003). Consciousness and its Place in Nature. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell Guide to Philosophy of Mind* (pp. 102-142). https://doi.org/10.1002/9780470998762.ch5

Chemero, A. (2009). *Radical embodied cognitive science*. The MIT Press.

Chiribella, G., & Maria Scandolo, C. (2015). Conservation of information and the foundations of quantum mechanics. *EPJ Web of Conferences*, *95*, 03003. https://doi.org/10.1051/epjconf/20149503003

Chisholm, R. (1964). Human Freedom and the Self. *The Lindley Lecture, Department of Philosophy, University of Kansas*.

Chisholm, R. M. (1976). The Agent as Cause. In M. Brand & D. Walton (Eds.), *Action Theory: Proceedings of the Winnipeg Conference on Human Action, Held at Winnipeg, Manitoba, Canada, 9–11 May 1975* (pp. 199-211). Springer Netherlands. https://doi.org/10.1007/978-94-010-9074-2_12

Chittka, L., Skorupski, P., & Raine, N. E. (2009). Speed–accuracy tradeoffs in animal decision making. *Trends in ecology & evolution*, *24*(7), 400-407. https://doi.org/10.1016/j.tree.2009.02.010

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67-90. https://doi.org/10.2307/2025900

Chvykov, P., Berrueta, T. A., Vardhan, A., Savoie, W., Samland, A., Murphey, T. D., Wiesenfeld, K., Goldman, D. I., & England, J. L. (2021). Low rattling: A predictive principle for self-organization in active collectives. *Science*, *371*(6524), 90-95. https://doi.org/10.1126/science.abc6182

Cisek, P. (1999). Beyond the computer metaphor: Behavior as interaction. *Journal of Consciousness Studies*, *6*(11-12), 125-142.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*, *81*(7), 2265-2287. https://doi.org/10.3758/s13414-019-01760-1

Cisek, P. (2022). Evolution of behavioural control from chordates to primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200522. https://doi.org/doi:10.1098/rstb.2020.0522

Cisek, P., & Kalaska, J. F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual review of neuroscience*, *33*(Volume 33, 2010), 269-298. https://doi.org/10.1146/annurev.neuro.051508.135409

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in Changing Conditions: The Urgency-Gating Model. *The Journal of Neuroscience*, *29*(37), 11560-11571. https://doi.org/10.1523/jneurosci.1844-09.2009

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001

Clarke, R. (1996). Agent Causation and Event Causation in the Production of Free Action. *Philosophical Topics*, *24*(2), 19-48.

Clarke, R. (2003). *Libertarian Accounts of Free Will*. Oxford University Press. https://doi.org/10.1093/019515987x.001.0001

Clarke, R. (2014). *Omissions: Agency, Metaphysics, and Responsibility*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199347520.001.0001

Clarke, R. (2019). Free Will, Agent Causation, and "Disappearing Agents". *Noûs*, *53*(1), 76-96. https://doi.org/10.1111/nous.12206

Clarke, R., Capes, J., & Swenson, P. (2021). Incompatibilist (Nondeterministic) Theories of Free Will. *The Stanford Encyclopedia of Philosophy (Fall 2021 Edition)*. URL = <https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/>

Cobb, M. (2020). *The Idea of the Brain: The Past and Future of Neuroscience*. Profile Books.

Comolatti, R., & Hoel, E. (2022). Causal emergence is widespread across measures of causation. *arXiv preprint*. https://doi.org/10.48550/arXiv.2202.01854

Conant, R. C., & Ashby, R. W. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*(2), 89-97. https://doi.org/10.1080/00207727008920220

Crick, F. H. C. (1994). *The Astonishing Hypothesis: The Scientific Search for the Soul*. Charles Scribner's Sons.

D'Ariano, G. M. (2018). Causality re-established. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2123), 20170313. https://doi.org/doi:10.1098/rsta.2017.0313

Danaher, J. (2019). The rise of the robots and the crisis of moral patiency. *AI & Society*, *34*(1), 129-136. https://doi.org/10.1007/s00146-017-0773-9

Dasgupta, S. (2016). *Computer Science: A Very Short Introduction*. Oxford University Press. https://doi.org/10.1093/actrade/9780198733461.001.0001

Dattathrani, S., & De', R. (2022). The Concept of Agency in the Era of Artificial Intelligence: Dimensions and Degrees. *Information Systems Frontiers*, *25*(1), 29–54. https://doi.org/10.1007/s10796-022-10336-8

Davidson, D. (1963). Actions, Reasons, and Causes. *The Journal of Philosophy*, *60*(23), 685-700. https://doi.org/10.2307/2023177

Dayan, P. (2012). Twenty-Five Lessons from Computational Neuromodulation. *Neuron*, *76*(1), 240-256. https://doi.org/10.1016/j.neuron.2012.09.027

de Calleja, M. P. (2014). Cross-world luck at the time of decision is a problem for compatibilists as well. *Philosophical Explorations*, *17*(2), 112-125.

Deacon, T. W. (2011). *Incomplete Nature: How mind emerged from matter*. WW Norton & Company.

Deco, G., & Kringelbach, M. L. (2020). Turbulent-like Dynamics in the Human Brain. *Cell reports*, *33*(10). https://doi.org/10.1016/j.celrep.2020.108471

Deco, G., Perl, Y. S., Jerotic, K., Escrichs, A., & Kringelbach, M. L. (2025). Turbulence as a framework for brain dynamics in health and disease. *Neuroscience & Biobehavioral Reviews*, *169*, 105988. https://doi.org/10.1016/j.neubiorev.2024.105988

Deco, G., & Rolls, E. T. (2006). Decision-making and Weber's law: a neurophysiological model. *European Journal of Neuroscience*, *24*(3), 901-916. https://doi.org/10.1111/j.1460-9568.2006.04940.x

Deco, G., Rolls, E. T., & Romo, R. (2009). Stochastic dynamics as a principle of brain function. *Progress in Neurobiology*, *88*(1), 1-16. https://doi.org/10.1016/j.pneurobio.2009.01.006

Del Santo, F. (2021). Indeterminism, Causality and Information: Has Physics Ever Been Deterministic? In A. Aguirre, Z. Merali, & D. Sloan (Eds.), *Undecidability, Uncomputability, and Unpredictability* (pp. 63-79). Springer International Publishing. https://doi.org/10.1007/978-3-030-70354-7_5

Del Santo, F., & Gisin, N. (2019). Physics without determinism: Alternative interpretations of classical physics. *Physical Review A*, *100*(6), 062107. https://doi.org/10.1103/PhysRevA.100.062107

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

Dennett, D. C. (2015). *Elbow Room: The Varieties of Free Will Worth Wanting*. NED-New edition, The MIT Press. https://doi.org/10.7551/mitpress/10470.001.0001

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*(4), 357-370. https://doi.org/10.1037/h0070405

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429-452.

Djedovic, A. (2020). *From Life-Like to Mind-Like Explanation: Natural Agency and the Cognitive Sciences* (Publication Number 27998040) [PhD thesis, University of Toronto].

Dohmatob, E., Dumas, G., & Bzdok, D. (2020). Dark control: The default mode network as a reinforcement learning agent. *Human Brain Mapping*, *41*(12), 3318-3341. https://doi.org/10.1002/hbm.25019

Doyle, B. (2011). *Free Will: The Scandal in Philosophy*.

Dresow, M., & Love, A. C. (2023). Teleonomy: Revisiting a Proposed Conceptual Replacement for Teleology. *Biological Theory*, *18*(2), 101-113. https://doi.org/10.1007/s13752-022-00424-y

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. The MIT Press.

Driscoll, L. N., Duncker, L., & Harvey, C. D. (2022). Representational drift: Emerging theories for continual learning and experimental future directions. *Current Opinion in Neurobiology*, *76*, 102609. https://doi.org/10.1016/j.conb.2022.102609

Drossel, B. (2015). On the Relation Between the Second Law of Thermodynamics and Classical and Quantum Mechanics. In B. Falkenburg & M. Morrison (Eds.), *Why More Is Different: Philosophical Issues in Condensed Matter Physics and Complex Systems* (pp. 41-54). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-43911-1_3

Drossel, B. (2017). Ten reasons why a thermalized system cannot be described by a many-particle wave function. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *58*, 12-21. https://doi.org/10.1016/j.shpsb.2017.04.001

Drossel, B. (2023). The passage of time and top-down causation. In *The Sixteenth Marcel Grossmann Meeting on Recent Developments in Theoretical and Experimental General Relativity, Astrophysics and Relativistic Field Theories* (pp. 3631-3645). https://doi.org/10.1142/9789811269776_0300

Drossel, B., & Ellis, G. (2018). Contextual Wavefunction collapse: an integrated theory of quantum measurement. *New Journal of Physics*, *20*(11), 113025. https://doi.org/10.1088/1367-2630/aaecec

Dupre, C., & Yuste, R. (2017). Non-overlapping Neural Networks in Hydra vulgaris. *Current Biology*, *27*(8), 1085-1097. https://doi.org/10.1016/j.cub.2017.02.049

Dupré, J. (2011). *Processes of Life: Essays in the Philosophy of Biology*. Oxford University Press.

Durstewitz, D., Huys, Q. J. M., & Koppe, G. (2021). Psychiatric Illnesses as Disorders of Network Dynamics. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 865-876. https://doi.org/10.1016/j.bpsc.2020.01.001

Durstewitz, D., Koppe, G., & Thurm, M. I. (2023). Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nature Reviews Neuroscience*, *24*(11), 693-710. https://doi.org/10.1038/s41583-023-00740-7

Earman, J. (2004). Determinism: What We Have Learned and What We Still Don't Know. In *Freedom and Determinism.* (pp. 21-46). Boston Review.

Earman, J. (2008). How Determinism Can Fail in Classical Physics and How Quantum Physics Can (Sometimes) Provide a Cure. *Philosophy of Science*, *75*(5), 817-829. https://doi.org/10.1086/594526

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. *Neuron*, *109*(19), 3055-3068. https://doi.org/10.1016/j.neuron.2021.07.011

Eddington, A. (1935). New Pathways in Science. *Philosophy*, *10*(40), 483-485.

Edmunds, L. (1972). Necessity, Chance, and Freedom in the Early Atomists. *Phoenix*, *26*(4), 342-357. https://doi.org/10.2307/1087594

Ekstrom, L. W. (1993). A coherence theory of autonomy. *Philosophy and Phenomenological Research*, *53*(3), 599-616.

Ekstrom, L. W. (1999). *Free Will: A Philosophical Study*. Westview.

Ekstrom, L. W. (2019). Toward a plausible event-causal indeterminist account of free will. *Synthese*, *196*(1), 127-144. https://doi.org/10.1007/s11229-016-1143-8

Ellis, G. F. R. (2008). On the nature of causation in complex systems. *Transactions of the Royal Society of South Africa*, *63*(1), 69-84. https://doi.org/10.1080/00359190809519211

Ellis, G. F. R. (2009). Top-down causation and the human brain. In N. Murphy, G. F. R. Ellis, & T. O'Connor (Eds.), *Downward causation and the neurobiology of free will* (pp. 63-81). Springer Verlag. https://doi.org/10.1007/978-3-642-03205-9_4

Ellis, G. F. R. (2012). Top-down causation and emergence: some comments on mechanisms. *Interface Focus*, *2*(1), 126-140. https://doi.org/doi:10.1098/rsfs.2011.0062

Ellis, G. F. R. (2016). *How Can Physics Underlie the Mind? Top-Down Causation in the Human Context*. Springer-Verlag.

Ellis, G. F. R. (2023). Quantum physics and biology: the local wavefunction approach. *Journal of Physics: Conference Series*, *2533*(1), 012019. https://doi.org/10.1088/1742-6596/2533/1/012019

Ellis, G. F. R., & Drossel, B. (2020). Emergence of Time. *Foundations of Physics*, *50*(3), 161-190.

Ellis, G. F. R., Meissner, K. A., & Nicolai, H. (2018). The physics of infinity. *Nature Physics*, *14*(8), 770-772. https://doi.org/10.1038/s41567-018-0238-1

England, J. L. (2022). Self-organized computation in the far-from-equilibrium cell. *Biophysics Reviews*, *3*(4). https://doi.org/10.1063/5.0103151

Erisken, S., Vaiceliunaite, A., Jurjut, O., Fiorini, M., Katzner, S., & Busse, L. (2014). Effects of Locomotion Extend throughout the Mouse Early Visual System. *Current Biology*, *24*(24), 2899-2907. https://doi.org/10.1016/j.cub.2014.10.045

Everett, H. (1957). "Relative State" Formulation of Quantum Mechanics. *Reviews of Modern Physics*, *29*(3), 454-462. https://doi.org/10.1103/RevModPhys.29.454

Eyink, G. L., & Drivas, T. D. (2015). Quantum spontaneous stochasticity. *arXiv preprint*. https://doi.org/10.48550/arXiv.1509.04941

Fábregas-Tejeda, A., & Baedke, J. (2023). Teleology, Organisms, and Genes: A Commentary on Haig. In T. E. Dickins & B. J. A. Dickins (Eds.), *Evolutionary Biology:*

Contemporary and Historical Reflections Upon Core Theory (pp. 249-264). Springer International Publishing. https://doi.org/10.1007/978-3-031-22028-9_15

Faisal, A. A., White, J. A., & Laughlin, S. B. (2005). Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring. *Current Biology*, *15*(12), 1143-1149. https://doi.org/10.1016/j.cub.2005.05.056

Falcon, A. (2023). Aristotle on Causality. *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. URL = <https://plato.stanford.edu/archives/spr2023/entries/aristotle-causality/>

Falke, J. J., Bass, R. B., Butler, S. L., Chervitz, S. A., & Danielson, M. A. (1997). The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. *Annual Review of Cell and Developmental Biology*, *13*, 457-512. https://doi.org/10.1146/annurev.cellbio.13.1.457

Farnsworth, K. D. (2018). How Organisms Gained Causal Independence and How It Might Be Quantified. *Biology (Basel)*, *7*(3). https://doi.org/10.3390/biology7030038

Farnsworth, K. D. (2022). How an information perspective helps overcome the challenge of biology to physics. *Biosystems*, *217*, 104683. https://doi.org/10.1016/j.biosystems.2022.104683

Félix, M.-A., & Barkoulas, M. (2015). Pervasive robustness in biological systems. *Nature Reviews Genetics*, *16*(8), 483-496. https://doi.org/10.1038/nrg3949

Filipowicz, A., Lalsiamthara, J., & Aballay, A. (2022). Dissection of a sensorimotor circuit underlying pathogen aversion in C. elegans. *BMC Biology*, *20*(1), 229. https://doi.org/10.1186/s12915-022-01424-x

Fischer, J. M. (2005). *Free Will: Critical Concepts in Philosophy* (Vol. 1). Routledge.

Fischer, J. M. (2007). Compatibilism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four Views on Free Will* (pp. 44-84). Blackwell.

Fischer, J. M. (2014). Toward a solution to the luck problem. In D. Palmer (Ed.), *Libertarian Free Will: Contemporary Debates*. Oxford University Press.

Fischer, J. M., Kane, R., Pereboom, D., & Vargas, M. (2007). *Four views on free will*. Blackwell Publishing.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511814594

Flack, J. C. (2017). Coarse-graining as a downward causation mechanism. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160338. https://doi.org/doi:10.1098/rsta.2016.0338

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280-1286. https://doi.org/doi:10.1098/rstb.2012.0021

Flynn, W. R. (1962). Visual hallucinations in sensory deprivation. *The Psychiatric Quarterly*, *36*(1), 55-65. https://doi.org/10.1007/BF01586100

Folse, H. J., 3rd, & Roughgarden, J. (2010). What is an individual organism? A multilevel selection perspective. *Q Rev Biol*, *85*(4), 447-472. https://doi.org/10.1086/656905

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, *66*(23), 829-839. https://doi.org/10.2307/2023833

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5-20. https://doi.org/10.2307/2024717

Franklin, C. E. (2013). How should libertarians conceive of the location and role of indeterminism? *Philosophical Explorations*, *16*(1), 44-58. https://doi.org/10.1080/13869795.2013.723036

Franklin, C. E. (2014). Event-causal libertarianism, functional reduction, and the disappearing agent argument. *Philosophical Studies*, *170*(3), 413-432. https://doi.org/10.1007/s11098-013-0237-0

Franklin, C. E. (2018). *A Minimal Libertarianism: Free Will and the Promise of Reduction*. Oxford University Press. https://doi.org/10.1093/oso/9780190682781.001.0001

Freddolino, L., & Tavazoie, S. (2012). Beyond Homeostasis: A Predictive-Dynamic Framework for Understanding Cellular Behavior. *Annual Review of Cell and Developmental Biology*, *28*, 363-384. https://doi.org/10.1146/annurev-cellbio-092910-154129

Freeman, W. J. (2000). A neurobiological interpretation of semiotics: meaning, representation, and information. *Information Sciences*, *124*(1), 93-102. https://doi.org/10.1016/S0020-0255(99)00144-9

Friston, K. (2005). A Theory of Cortical Responses. *Philosophical Transactions: Biological Sciences*, *360*(1456), 815-836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787

Friston, K. (2018). Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?) [Hypothesis and Theory]. *Frontiers in Psychology*, *Volume 9 - 2018*. https://doi.org/10.3389/fpsyg.2018.00579

Froese, T. (2023). Irruption Theory: A Novel Conceptualization of the Enactive Account of Motivated Activity. *Entropy (Basel)*, *25*(5). https://doi.org/10.3390/e25050748

Froese, T., & Taguchi, S. (2019). The Problem of Meaning in AI and Robotics: Still with Us after All These Years. *Philosophies*, *4*(2), 14.

Fulda, F. (2016). *Natural Agency: An Ecological Approach.* [PhD thesis, University of Toronto].

Fulda, F. (2017). Natural Agency: The Case of Bacterial Cognition. *Journal of the American Philosophical Association*, *3*(1), 69-90. https://doi.org/10.1017/apa.2017.5

Fulda, F. (2023). Agential autonomy and biological individuality. *Evolution & Development*, *25*(6), 353-370. https://doi.org/10.1111/ede.12450

Gallego, J. A., Perich, M. G., Miller, L. E., & Solla, S. A. (2017). Neural Manifolds for the Control of Movement. *Neuron*, *94*(5), 978-984. https://doi.org/10.1016/j.neuron.2017.05.025

García-Valdecasas, M., & Deacon, T. W. (2024). Origins of Biological Teleology: How Constraints Represent Ends. *Synthese*, *204*(75), 1-28.

Genkin, M., Shenoy, K. V., Chandrasekaran, C., & Engel, T. A. (2025). The dynamics and geometry of choice in the premotor cortex. *Nature*. https://doi.org/10.1038/s41586-025-09199-1

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287-292. https://doi.org/10.1016/S1364-6613(03)00128-1

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165-193. https://doi.org/10.1016/0010-0277(95)00661-H

Ghalambor, C. K., Hoke, K. L., Ruell, E. W., Fischer, E. K., Reznick, D. N., & Hughes, K. A. (2015). Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature*, *525*(7569), 372-375.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton, Mifflin and Company.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual review of psychology*, *62*, 451-482. https://doi.org/10.1146/annurev-psych-120709-145346

Ginzburg, V. L., & Syrovatskii, S. I. (1964). *The Origin of Cosmic Rays* (H. S. H. Massey, Trans.; D. Ter Haar, Ed.). Pergamon Press.

Gisin, N. (2021a). Indeterminism in physics and intuitionistic mathematics. *Synthese*, *199*(5), 13345-13371. https://doi.org/10.1007/s11229-021-03378-z

Gisin, N. (2021b). Indeterminism in Physics, Classical Chaos and Bohmian Mechanics: Are Real Numbers Really Real? *Erkenntnis*, *86*(6), 1469-1481. https://doi.org/10.1007/s10670-019-00165-8

Glasscock, J. P., & Tenenbaum, S. (2023). “Action”. *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. URL = <https://plato.stanford.edu/archives/spr2023/entries/action/>

Glimcher, P. W. (2003). *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. The MIT Press. https://doi.org/10.7551/mitpress/2302.001.0001

Glimcher, P. W. (2005). Indeterminacy in Brain and Behavior. *Annual review of psychology*, *56*, 25-56. https://doi.org/10.1146/annurev.psych.55.090902.141429

Glimcher, P. W., & Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. *Science*, *306*(5695), 447-452.

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual review of neuroscience*, *30*, 535-574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goldman, A. I. (1970). *Theory of Human Action*. Princeton University Press.

Gomez-Marin, A. (2017). Causal Circuit Explanations of Behavior: Are Necessity and Sufficiency Necessary and Sufficient? In A. Çelik & M. F. Wernet (Eds.), *Decoding Neural Circuit Structure and Function: Cellular Dissection Using Genetic Model Organisms* (pp. 283-306). Springer International Publishing. https://doi.org/10.1007/978-3-319-57363-2_11

González-Rueda, A., Jensen, K., Noormandipour, M., de Malmazet, D., Wilson, J., Ciabatti, E., Kim, J., Williams, E., Poort, J., Hennequin, G., & Tripodi, M. (2024). Kinetic features dictate sensorimotor alignment in the superior colliculus. *Nature*, *631*(8020), 378-385. https://doi.org/10.1038/s41586-024-07619-2

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085-1108. https://doi.org/10.1037/a0028044

Gottlieb, G. (1976). Conceptions of prenatal development: Behavioral embryology. *Psychological Review*, *83*(3), 215-234. https://doi.org/10.1037/0033-295X.83.3.215

Greene, B. (2021). *Until the end of time: Mind, matter, and our search for meaning in an evolving universe*. Vintage.

Griffith, M. (2010). Why agent-caused actions are not lucky. *American Philosophical Quarterly*, *47*(1), 43-56.

Grillner, S., & Robertson, B. (2016). The Basal Ganglia Over 500 Million Years. *Current Biology*, *26*(20), R1088-R1100. https://doi.org/10.1016/j.cub.2016.06.041

Guertin, P. A. (2012). Central Pattern Generator for Locomotion: Anatomical, Physiological, and Pathophysiological Considerations [Review]. *Frontiers in Neurology*, *3,* 183. https://doi.org/10.3389/fneur.2012.00183

Haji, I. (2004). Active control, agent-causation and free action. *Philosophical Explorations*, *7*(2), 131-148. https://doi.org/10.1080/13869790410001694480

Hall, N. (2004). Two Concepts of Causation. In J. Collins, N. Hall, & L. Paul (Eds.), *Causation and Counterfactuals* (pp. 225-276). The MIT Press.

Hallgrimsson, B., Green, R. M., Katz, D. C., Fish, J. L., Bernier, F. P., Roseman, C. C., Young, N. M., Cheverud, J. M., & Marcucio, R. S. (2019). The developmental-genetics of canalization. *Seminars in Cell & Developmental Biology*, *88*, 67-79. https://doi.org/10.1016/j.semcdb.2018.05.019

Harris, S. (2012). *Free Will*. Free Press.

Hawking, S., & Mlodinow, L. (2010). *The Grand Design: New Answers to the Ultimate Questions of Life*. Bantam Press.

Hawking, S. W., & Ellis, G. F. R. (2024). *The Large Scale Structure of Space Time, 50th Anniversary Edition.* Cambridge University Press.

Hegel, G. W. F. (1807/1977). *Phenomenology of spirit* (A. V. Miller, Trans.). Clarendon Press.

Hesse, J., & Gross, T. (2014). Self-organized criticality as a fundamental property of neural systems [Review]. *Frontiers in Systems Neuroscience*, *Volume 8 - 2014*. https://doi.org/10.3389/fnsys.2014.00166

Himmelreich, J. (2019). Responsibility for Killer Robots. *Ethical Theory and Moral Practice*, *22*(3), 731-747. https://doi.org/10.1007/s10677-019-10007-9

Hobbes, T. (1651). *Leviathan* (Project Gutenberg, Trans.).

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500-544. https://doi.org/10.1113/jphysiol.1952.sp004764

Hoel, E. P. (2017). When the Map Is Better Than the Territory. *Entropy*, *19*(5), 188.

Hoel, E. P. (2018). Agent Above, Atom Below: How Agents Causally Emerge from Their Underlying Microphysics. In A. Aguirre, B. Foster, & Z. Merali (Eds.), *Wandering Towards a Goal: How Can Mindless Mathematical Laws Give Rise to Aims and Intention?* (pp. 63-76). Springer International Publishing. https://doi.org/10.1007/978-3-319-75726-1_6

Hoel, E. P. (2025). Causal Emergence 2.0: Quantifying emergent complexity. *arXiv preprint*. https://doi.org/10.48550/arXiv.2503.13395

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, *110*(49), 19790-19795. https://doi.org/doi:10.1073/pnas.1314922110

Honderich, T. (2001). Determinism as true, both compatibilism and incompatibilism as false, and the real problem. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 461-476). Oxford University Press.

Honderich, T. (2002). *How Free Are You? The Determinism Problem* (Second Edition). Oxford University Press.

Hooker, C. (2013). On the Import of Constraints in Complex Dynamical Systems. *Foundations of Science*, *18*(4), 757-780. https://doi.org/10.1007/s10699-012-9304-9

Hornsby, J. (2004). Agency and actions. *Royal Institute of Philosophy Supplements*, *55*, 1-23.

Hossenfelder, S. (2022). *Existential physics: A scientist's guide to life's biggest questions*. Penguin.

Hu, Y., Linz, D. M., Parker, E. S., Schwab, D. B., Casasa, S., Macagno, A. L. M., & Moczek, A. P. (2020). Developmental bias in horned dung beetles and its contributions to innovation, adaptation, and resilience. *Evolution & Development*, *22*(1-2), 165-180. https://doi.org/10.1111/ede.12310

Ismael, J. T. (2011). Self-Organization and Self-Governance. *Philosophy of the Social Sciences*, *41*(3), 327-351. https://doi.org/10.1177/0048393110363435

Ismael, J. T. (2016). *How Physics Makes Us Free*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190269449.001.0001

Jaeger, J. (2024). Ontogenesis, Organisation, and Organismal Agency. In J. Švorcová (Ed.), *Organismal Agency: Biological Concepts and Their Philosophical Foundations* (pp. 165-190). Springer International Publishing. https://doi.org/10.1007/978-3-031-53626-7_10

Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., & Walsh, D. (2024). Naturalizing relevance realization: why agency and cognition are fundamentally not computational [Hypothesis and Theory]. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1362658

James, W. (1884). *The dilemma of determinism*. Kessinger Publishing.

James, W. (1890). *The principles of psychology, Vol I*. Henry Holt and Co. https://doi.org/10.1037/10538-000

Jékely, G., Godfrey-Smith, P., & Keijzer, F. (2021). Reafference and the origin of the self in early nervous system evolution. *Philosophical Transactions of the Royal Society B:*

Biological Sciences, *376*(1821), 20190764. https://doi.org/doi:10.1098/rstb.2019.0764

Johnson, D. G., & Verdicchio, M. (2019). AI, agency and responsibility: the VW fraud case and beyond. *AI & Society*, *34*(3), 639-647. https://doi.org/10.1007/s00146-017-0781-9

Joslyn, C. (2000). Levels of Control and Closure in Complex Semiotic Systems. *Annals of the New York Academy of Sciences*, *901*(1), 67-74. https://doi.org/10.1111/j.1749-6632.2000.tb06266.x

Juarrero, A. (1999). *Dynamics in Action: Intentional Behavior as a Complex System*. The MIT Press. https://doi.org/10.7551/mitpress/2528.001.0001

Juarrero, A. (2015). What does the closure of context-sensitive constraints mean for determinism, autonomy, self-determination, and agency? *Progress in Biophysics and Molecular Biology*, *119*(3), 510-521. https://doi.org/10.1016/j.pbiomolbio.2015.08.007

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

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica*, *47*, 163-291.

Kane, R. (1996). *The Significance of Free Will*. Oxford University Press.

Kane, R. (1999). Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *The Journal of Philosophy*, *96*(5), 217-240. https://doi.org/10.2307/2564666

Kane, R. (2001). *The Oxford Handbook of Free Will* (Vol. 55). Oxford University Press.

Kane, R. (2007). Libertarianism. In J. M. Fischer, R. Kane, D. Pereboom, & M. Vargas (Eds.), *Four Views on Free Will*. Blackwell.

Kane, R. (2016). On the role of indeterminism in libertarian free will. *Philosophical Explorations*, *19*(1), 2-16. https://doi.org/10.1080/13869795.2016.1085594

Kane, R. (2019). The complex tapestry of free will: striving will, indeterminism and volitional streams. *Synthese*, *196*(1), 145-160. https://doi.org/10.1007/s11229-016-1046-8

Kane, R. (2024). *The Complex Tapestry of Free Will: A Philosophical Odyssey*. Oxford University Press. https://doi.org/10.1093/oso/9780197751404.001.0001

Kant, I. (1781/1929). *Critique of Pure Reason* (N. K. Smith, Trans.). MacMillan and Co.

Kant, I. (1788/2002). *Critique of Practical Reason* (W. S. Pluhar, Trans.). Hackett Publishing.

Kaplan, H. S., & Zimmer, M. (2020). Brain-wide representations of ongoing behavior: a universal principle? *Current Opinion in Neurobiology*, *64*, 60-69. https://doi.org/10.1016/j.conb.2020.02.008

Kastner, R. E. (2017). On Quantum Collapse as a Basis for the Second Law of Thermodynamics. *Entropy*, *19*(3), 106.

Kauffman, S. (2000). *Investigations*. Oxford University Press. https://doi.org/10.1093/oso/9780195121049.001.0001

Kauffman, S. (2013). What Is Life, and Can We Create It? *BioScience*, *63*(8), 609-610. https://doi.org/10.1525/bio.2013.63.8.2

Keijzer, F., van Duijn, M., & Lyon, P. (2013). What nervous systems do: early evolution, input–output, and the skin brain thesis. *Adaptive Behavior*, *21*(2), 67-85. https://doi.org/10.1177/1059712312465330

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461-468. https://doi.org/10.1016/S1364-6613(99)01402-3

Khilkevich, A., Lohse, M., Low, R., Orsolic, I., Bozic, T., Windmill, P., & Mrsic-Flogel, T. D. (2024). Brain-wide dynamics linking sensation to action during decision-making. *Nature*, *634*(8035), 890-900. https://doi.org/10.1038/s41586-024-07908-w

Kim, C. K., Adhikari, A., & Deisseroth, K. (2017). Integration of optogenetics with complementary methodologies in systems neuroscience. *Nature Reviews Neuroscience*, *18*(4), 222-235. https://doi.org/10.1038/nrn.2017.15

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press.

Kim, J. (1998). *Mind in a physical world: An essay on the mind–body problem and mental causation* (Vol. 75). The MIT Press.

Kim, J. (2000). Making sense of downward causation. In P. B. Andersen, C. Emmeche, N. O. Finnemann, & P. V. Christiansen (Eds.), *Downward Causation* (pp. 305-321). University of Aarhus Press.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kim, J. (2006). Emergence: Core ideas and issues. *Synthese*, *151*(3), 547-559. https://doi.org/10.1007/s11229-006-9025-0

Klein, B., & Hoel, E. (2020). The Emergence of Informative Higher Scales in Complex Networks. *Complexity*, *2020*, 8932526. https://doi.org/10.1155/2020/8932526

Klein, J. (2020). Francis Bacon. *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*. URL = <https://plato.stanford.edu/archives/fall2020/entries/francis-bacon/>

Krakauer, D., Bertschinger, N., Olbrich, E., Flack, J. C., & Ay, N. (2020). The information theory of individuality. *Theory in Biosciences*, *139*(2), 209-223. https://doi.org/10.1007/s12064-020-00313-7

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480-490. https://doi.org/10.1016/j.neuron.2016.12.041

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480-498. https://doi.org/10.1037/0033-2909.108.3.480

Kwok, W. (2020). Should Classical Physics Be Interpreted Indeterministically? *arXiv preprint*. https://doi.org/10.48550/arXiv.2005.07079

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, E., & Odling-Smee, J. (2015). The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings of the royal society B: biological sciences*, *282*(1813), 20151019.

Landauer, R. (1991). Information is Physical. *Physics Today*, *44*(5), 23-29. https://doi.org/10.1063/1.881299

Laplace, P. S. (1814). *Essai philosophique sur les probabilités*. Courcier.

Laskar, J. (1990). The chaotic motion of the solar system: A numerical estimate of the size of the chaotic zones. *Icarus*, *88*(2), 266-291. https://doi.org/10.1016/0019-1035(90)90084-M

Laskar, J. (2013). Is the Solar System Stable? In B. Duplantier, S. Nonnenmacher, & V. Rivasseau (Eds.), *Chaos: Poincaré Seminar 2010* (pp. 239-270). Springer Basel. https://doi.org/10.1007/978-3-0348-0697-8_7

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369-381.

Law, A. (2023). Incompatibilism and the garden of forking paths. *Philosophical Issues*, *33*(1), 110-123. https://doi.org/10.1111/phis.12247

Layzer, D. (1975). The Arrow of Time. *Scientific American*, *233*(6), 56-69. https://doi.org/10.1038/scientificamerican1275-56

Layzer, D. (2021). *Why we are free: Consciousness, free will and creativity in a unified scientific worldview*. I-Phi Press.

Lea, A. J., Tung, J., Archie, E. A., & Alberts, S. C. (2018). Developmental plasticity: Bridging research in evolution and human health. *Evolution, medicine, and public health*, *2017*(1), 162-175. https://doi.org/10.1093/emph/eox019

Lee, S. Y., Kozalakis, K., Baftizadeh, F., Campagnola, L., Jarsky, T., Koch, C., & Anastassiou, C. A. (2024). Cell-class-specific electric field entrainment of neural activity. *Neuron*, *112*(15), 2614-2630.e2615. https://doi.org/10.1016/j.neuron.2024.05.009

Leffler, O. (2025). Even Bigger-Picture Causalism. *Acta Analytica*. https://doi.org/10.1007/s12136-025-00647-1

LemaÎtre, G. (1931). The Beginning of the World from the Point of View of Quantum Theory. *Nature*, *127*(3210), 706-706. https://doi.org/10.1038/127706b0

Lemos, J. (2021). The Indeterministic Weightings Model of Libertarian Free Will. *Journal of Philosophical Theological Research*, *23*(3), 137-156.

Levin, M. (2022). Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds [Hypothesis and Theory]. *Frontiers in Systems Neuroscience*, *16*. https://doi.org/10.3389/fnsys.2022.768201

Levin, M., & Dennett, D. C. (2020). Cognition all the way down. *Aeon Essays*.

Levy, N. (2008). Bad luck once again. *Philosophy and Phenomenological Research*, *77*(3), 749-754.

Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford University Press.

Levy, N. (2014). *Consciousness and Moral Responsibility*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198704638.001.0001

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*, *6*(8-9), 47-57.

Libet, B., Wright, E. W., & Gleason, C. A. (1982). Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. *Electroencephalography and Clinical Neurophysiology*, *54*(3), 322-335. https://doi.org/10.1016/0013-4694(82)90181-X

Lignani, G., Baldelli, P., & Marra, V. (2020). Homeostatic Plasticity in Epilepsy [Perspective]. *Frontiers in Cellular Neuroscience*, *14*. https://doi.org/10.3389/fncel.2020.00197

List, C. (2019). *Why Free Will Is Real*. Harvard University Press.

List, C. (2021). Group Agency and Artificial Intelligence. *Philosophy & Technology*, *34*(4), 1213-1242. https://doi.org/10.1007/s13347-021-00454-7

List, C., & Menzies, P. (2017). My brain made me do it: The exclusion argument against free will, and what's wrong with it. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a Difference: Essays on the Philosophy of Causation*. Oxford University Press.

List, C., & Rabinowicz, W. (2014). Two Intuitions About Free Will: Alternative Possibilities And Intentional Endorsement. *Philosophical Perspectives*, *28*, 155-172.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, *20*(2), 130-141.

Lorenz, E. N. (1969). The predictability of a flow which possesses many scales of motion. *Tellus*, *21*(3), 289-307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

Lorimer, K., Knight, R., & Shoveller, J. (2022). Improving the health and social wellbeing of young people: exploring the potential of and for collective agency. *Critical Public Health*, *32*(2), 145-152. https://doi.org/10.1080/09581596.2020.1786501

Louis, K. S., & Khalifa, M. (2018). Understanding and improving urban secondary schools: the role of individual and collective agency. *Journal of Educational Administration*, *56*(5), 446-454. https://doi.org/10.1108/JEA-08-2018-174

Lydic, R. (1989). Central pattern-generating neurons and the search for general principles. *The FASEB Journal*, *3*(13), 2457-2468. https://doi.org/10.1096/fasebj.3.13.2680703

Lyon, P. (2015). The cognitive cell: bacterial behavior reconsidered [Review]. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00264

Mackie, P. (2018). Compatibilism, Indeterminism, and Chance. *Royal Institute of Philosophy Supplement*, *82*, 265-287. https://doi.org/10.1017/S1358246118000140

Malpas, J. (2024). "Donald Davidson". *The Stanford Encyclopedia of Philosophy (Fall 2024 Edition)*. URL = <https://plato.stanford.edu/archives/fall2024/entries/davidson/>

Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. *eLife*, *8*, e39787. https://doi.org/10.7554/eLife.39787

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175-183. https://doi.org/10.1007/s10676-004-3422-1

Mariani, C., & Torrengo, G. (2021). The indeterminate present and the open future. *Synthese*, *199*(1/2), 3923-3944.

Markosian, N. (1999). A Compatibilist Version Of The Theory Of Agent Causation. *Pacific Philosophical Quarterly*, *80*(3), 257-277. https://doi.org/10.1111/1468-0114.00083

Markosian, N. (2012). Agent causation as the solution to all the compatibilist's problems. *Philosophical Studies*, *157*(3), 383-398.

Marr, D. C., & Poggio, T. A. (1976). From Understanding Computation to Understanding Neural Circuitry. MIT Artificial Intelligence Laboratory,

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living* (1st Edition). Springer Dordrecht. https://doi.org/10.1007/978-94-009-8947-4

Mayr, E. (1974). Teleological and Teleonomic, a New Analysis. In R. S. Cohen & M. W. Wartofsky (Eds.), *A Portrait of Twenty-five Years: Boston Colloquium for the Philosophy of Science 1960–1985* (pp. 133-159). Springer Netherlands. https://doi.org/10.1007/978-94-009-5345-1_10

Mayr, E. (1988). *Toward a new philosophy of biology: observations of an evolutionist*. Belknap Press of Harvard University Press.

McCall, S., & Lowe, E. J. (2005). Indeterminist Free Will. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/j.1933-1592.2005.tb00420.x

McCall, S., & Lowe, E. J. (2008). The Determinists Have Run Out of Luck: For a Good Reason. *Philosophy and Phenomenological Research*. https://doi.org/10.1111/j.1933-1592.2008.00218.x

McCormick, D. A., Nestvogel, D. B., & He, B. J. (2020). Neuromodulation of Brain State and Behavior. *Annual review of neuroscience*, *43*, 391-415. https://doi.org/10.1146/annurev-neuro-100219-105424

McKenna, M. (2003). Robustness, control, and the demand for morally significant alternatives: Frankfurt examples with oodles and oodles of alternatives. In M. S. McKenna & D. Widerker (Eds.), *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (pp. 201--217). Ashgate.

McKenna, M., & Coates, J. D. (2024). "Compatibilism". *The Stanford Encyclopedia of Philosophy (Fall 2024 Edition)*. URL = <https://plato.stanford.edu/archives/fall2024/entries/compatibilism/>

McKenna, M., & Pereboom, D. (2016). *Free Will: A Contemporary Introduction*. Routledge.

Meena, C., Hens, C., Acharyya, S., Haber, S., Boccaletti, S., & Barzel, B. (2023). Emergent stability in complex network dynamics. *Nature Physics*, *19*(7), 1033-1042. https://doi.org/10.1038/s41567-023-02020-8

Meincke, A. S. (2018). Autopoiesis, biological autonomy and the process view of life. *European Journal for Philosophy of Science*, *9*(1), 1-16.

Mele, A. R. (2005). Libertarianism, luck, and control. *Pacific Philosophical Quarterly*, *86*(3), 381-407.

Mele, A. R. (2006). *Free Will and Luck*. Oxford University Press. https://doi.org/10.1093/0195305043.001.0001

Mele, A. R. (2024a). On a Disappearing Agent Argument: Settling Matters. *The Journal of Ethics*, *28*(2), 351-360. https://doi.org/10.1007/s10892-023-09438-5

Mele, A. R. (2024b). Libertarianism, decision-making, and a point of no return. *Philosophical Studies*, *181*(9), 2391-2404. https://doi.org/10.1007/s11098-024-02190-y

Menzies, P., & List, C. (2010). The Causal Autonomy of the Special Sciences. In G. Macdonald & C. Macdonald (Eds.), *Emergence in mind* (pp. 108-129). Oxford University Press.

Merleau-Ponty, M. (1945/1970). *Phenomenology of perception* (C. Smith, Trans.). Routledge & Kegan Paul.

Mestek-Boukhibar, L., & Barkoulas, M. (2015). The developmental genetics of biological robustness. *Annals of Botany*, *117*(5), 699-707. https://doi.org/10.1093/aob/mcv128

Meyer, S. S. (2014). Aristotle on what is up to us and what is contingent. In P. S. R. Destrée & M. Zingano (Eds.), *What is Up to Us? Studies on Agency and Responsibility in ancient Philosophy.* Sankt Augustin: Academia Verlag.

Michod, R. E. (2007). Evolution of individuality during the transition from unicellular to multicellular life. *Proceedings of the National Academy of Sciences*, *104*(suppl 1), 8613-8618. https://doi.org/10.1073/pnas.0701489104

Miller, P. (2016). Dynamical systems, attractors, and neural circuits. *F1000Research*, *5*(992). https://doi.org/10.12688/f1000research.7698.1

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations for Realism*. The MIT Press. https://doi.org/10.7551/mitpress/4124.001.0001

Millikan, R. G. (1995). Pushmi-Pullyu Representations. *Philosophical Perspectives*, *9*, 185-200. https://doi.org/10.2307/2214217

Milonni, P. W. (1994). *The Quantum Vacuum: An Introduction to Quantum Electrodynamics*. Academic Press. https://doi.org/10.1016/C2009-0-21295-5

Mitchell, K. J. (2015). The Genetic Architecture of Neurodevelopmental Disorders. In K. J. Mitchell (Ed.), *The genetics of neurodevelopmental disorders* (pp. 1-28). https://doi.org/10.1002/9781118524947.ch1

Mitchell, K. J. (2018a). *Innate: How the wiring of our brains shapes who we are*. Princeton University Press.

Mitchell, K. J. (2018b). Does Neuroscience Leave Room for Free Will? *Trends in Neurosciences*, *41*(9), 573-576. https://doi.org/10.1016/j.tins.2018.05.008

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

Mitchell, K. J. (2023b). The origins of meaning – from pragmatic control signals to semantic representations. *OSF Preprint*. https://doi.org/10.31234/osf.io/dfkrv

Mitchell, K. J. (2025). Undetermined: Free will in real time and through time. *Teorema. International Journal of Philosophy*, *44*(1). https://doi.org/10.30827/trif.34308

Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, *372*, 179-191. https://doi.org/10.1016/j.jtbi.2015.02.029

Moore, D. (2021). Libertarian Free Will and the Physical Indeterminism Luck Objection. *Philosophia*, *50*(1), 159-182.

Moore, D. (2023). Lemos on the Physical Indeterminism Luck Objection. *Philosophia*, *51*(3), 1459-1477. https://doi.org/10.1007/s11406-022-00591-z

Moreno, A., & Mossio, M. (2015). *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Springer Dordrecht. https://doi.org/10.1007/978-94-017-9837-2

Moritz, S. E., Feltz, D. L., Fahrbach, K. R., & Mack, D. E. (2000). The relation of self-efficacy measures to sport performance: A meta-analytic review. *Research Quarterly for Exercise and Sport*, *71*(3), 280-294. https://doi.org/10.1080/02701367.2000.10608908

Morse, S. J. (1994). Culpability and Control. *University of Pennsylvania Law Review*, *142*(5), 1587-1660. https://doi.org/10.2307/3312464

Müller, J.-F. (n.d.). A New Role for Rollbacks: Showing How Objective Probabilities Undermine the Ability to Act Otherwise. *PhilArchive*. https://philarchive.org/rec/JANANR-2

Müller, T., & Briegel, H. J. (2018). A Stochastic Process Model for Free Agency under Indeterminism. *Dialectica*, *72*(2), 219-252. https://doi.org/10.1111/1746-8361.12222

Müller, T., Rumberg, A., & Wagner, V. (2019). An introduction to real possibilities, indeterminism, and free will: three contingencies of the debate. *Synthese*, *196*(1), 1-10. https://doi.org/10.1007/s11229-018-1842-4

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology*, *38*(1), 30-38. https://doi.org/10.1037/0022-0167.38.1.30

Mumford, S., & Anjum, R. L. (2015). Freedom and control: on the modality of free will. *American Philosophical Quarterly*, 1-11.

Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S., & Churchland, A. K. (2019). Single-trial neural dynamics are dominated by richly varied movements. *Nature Neuroscience*, *22*(10), 1677-1686. https://doi.org/10.1038/s41593-019-0502-4

Nadolski, E. M., & Moczek, A. P. (2023). Promises and limits of an agency perspective in evolutionary developmental biology. *Evolution & Development*, *25*(6), 371-392. https://doi.org/10.1111/ede.12432

Nave, K., Deane, G., Miller, M., & Clark, A. (2022). Expecting some action: Predictive Processing and the construction of conscious experience. *Review of Philosophy and Psychology*, *13*(4), 1019-1037. https://doi.org/10.1007/s13164-022-00644-y

Neuberger, E. J., Gupta, A., Subramanian, D., Korgaonkar, A. A., & Santhakumar, V. (2019). Converging early responses to brain injury pave the road to epileptogenesis. *Journal of neuroscience research*, *97*(11), 1335-1344. https://doi.org/10.1002/jnr.24202

Newman, S. A. (2023). Form, Function, Agency: Sources of Natural Purpose in Animal Evolution. In P. A. Corning, S. A. Kauffman, D. Noble, J. A. Shapiro, R. I. Vane-Wright, & A. Pross (Eds.), *Evolution "On Purpose": Teleonomy in Living Systems*. The MIT Press. https://doi.org/10.7551/mitpress/14642.003.0014

Newton, I. (1730/1952). *Opticks* (Fourth Edition). Dover Publications.

Neyrinck, M., Genel, S., & Stücker, J. (2022). Boundaries of chaos and determinism in the cosmos. *arXiv preprint.* https://doi.org/10.48550/arXiv.2206.10666

Nicholson, D. J. (2019). Is the cell really a machine? *Journal of Theoretical Biology*, *477*, 108-126. https://doi.org/10.1016/j.jtbi.2019.06.002

Nicholson, D. J., & Dupré, J. (2018). *Everything Flows: Towards a Processual Philosophy of Biology*. Oxford University Press. https://doi.org/10.1093/oso/9780198779636.001.0001

Nietzsche, F. (1886). *Beyond Good and Evil* (H. Zimmern, Trans.). The Modern Library.

Nijhout, H. F., Sadre-Marandi, F., Best, J., & Reed, M. C. (2017). Systems Biology of Phenotypic Robustness and Plasticity. *Integrative and Comparative Biology*, *57*(2), 171-184. https://doi.org/10.1093/icb/icx076

Nozick, R. (1981). *Philosophical explanations*. Harvard University Press.

O'Connell, R. G., & Kelly, S. P. (2021). Neurophysiology of human perceptual decision-making. *Annual review of neuroscience*, *44*(1), 495-516.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press.

O'Connor, T. (2009). Agent-causal power. In T. Handfield (Ed.), *Dispositions and Causes*. Oxford University Press.

O'Connor, T., & Franklin, C. E. (2022). "Free Will". *The Stanford Encyclopedia of Philosophy (Winter 2022 Edition)*. URL = <https://plato.stanford.edu/archives/win2022/entries/freewill/>

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 883-917.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Okasha, S. (2023). The Concept of Agent in Biology: Motivations and Meanings. *Biological Theory*, *19*(1), 6-10. https://doi.org/10.1007/s13752-023-00439-z

Oyama, S. (2000). *The ontogeny of information: Developmental systems and evolution.* (2nd Edition). Duke University Press. https://doi.org/10.1215/9780822380665

Palmer, D. (2014). *Libertarian Free Will: Contemporary Debates*. Oxford University Press.

Palmer, T. N., Döring, A., & Seregin, G. (2014). The real butterfly effect. *Nonlinearity*, *27*(9), R123. https://doi.org/10.1088/0951-7715/27/9/R123

Papineau, D. (2002). *Thinking about Consciousness*. Oxford University Press. https://doi.org/10.1093/0199243824.001.0001

Papineau, D. (2023). "Naturalism". *The Stanford Encyclopedia of Philosophy (Fall 2023 Edition)*. URL = <https://plato.stanford.edu/archives/fall2023/entries/naturalism/>

Payne, J. L., Moore, J. H., & Wagner, A. (2014). Robustness, Evolvability, and the Logic of Genetic Regulation. *Artificial Life*, *20*(1), 111-126. https://doi.org/10.1162/ARTL_a_00099

Pearl, J. (2009). *Causality* (Second Edition). Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511803161

Percival, I. A. N. (1991). Schrödinger's quantum cat. *Nature*, *351*(6325), 357-357. https://doi.org/10.1038/351357a0

Pereboom, D. (2004). Is Our Conception of Agent-Causation Coherent? *Philosophical Topics*, *32*(1/2), 275-286.

Pereboom, D. (2005). Defending Hard Incompatibilism. *Midwest Studies In Philosophy*, *29*(1), 228-247. https://doi.org/10.1111/j.1475-4975.2005.00114.x

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685516.001.0001

Pereboom, D. (2017). Responsibility, Agency, and the Disappearing Agent Objection. In *Le libre arbitre*. Collège de France. https://doi.org/10.4000/books.cdf.4942

Pereboom, D. (2022). *Free Will*. Cambridge University Press. https://doi.org/10.1017/9781108982511

Pérez-Ortega, J., Akrouh, A., & Yuste, R. (2024). Stimulus encoding by specific inactivation of cortical neurons. *Nature Communications*, *15*(1), 3192. https://doi.org/10.1038/s41467-024-47515-x

Pessoa, L. (2022). *The Entangled Brain: How perception, cognition, and emotion are woven together*. The MIT Press.

Pezzulo, G., & Castelfranchi, C. (2007). The symbol detachment problem. *Cognitive Processing*, *8*(2), 115-131. https://doi.org/10.1007/s10339-007-0164-0

Pezzulo, G., & Nolfi, S. (2019). Making the Environment an Informative Place: A Conceptual Analysis of Epistemic Policies and Sensorimotor Coordination. *Entropy*, *21*(4), 350.

Piccinini, G. (2022). Situated Neural Representations: Solving the Problems of Content [Hypothesis and Theory]. *Frontiers in Neurorobotics*, *16*. https://doi.org/10.3389/fnbot.2022.846979

Pickering, A. (2024). What Is Agency? A View from Science Studies and Cybernetics. *Biological Theory*, *19*(1), 16-21. https://doi.org/10.1007/s13752-023-00437-1

Pigliucci, M. (2005). Evolution of phenotypic plasticity: where are we going now? *Trends in ecology & evolution*, *20*(9), 481-486. https://doi.org/10.1016/j.tree.2005.06.001

Pinotsis, D. A., Fridman, G., & Miller, E. K. (2023). Cytoelectric coupling: Electric fields sculpt neural activity and "tune" the brain's infrastructure. *Progress in Neurobiology*, *226*, 102465. https://doi.org/10.1016/j.pneurobio.2023.102465

Piquero, A. R. (2021). Reflections on Choice and Agency in Context: a Reply. *Journal of Developmental and Life-Course Criminology*, *7*(4), 711-721. https://doi.org/10.1007/s40865-021-00174-8

Pittendrigh, C. (1958). Adaptation, natural selection, and behavior. In Roe A & S. GG (Eds.), *Behavior and evolution* (pp. 390–416). Yale University Press.

Pleasants, N. (2010). Moral Argument Is Not Enough: The Persistence of Slavery and the Emergence of Abolition. *Philosophical Topics*, *38*(1), 159-180.

Poincaré, H. (1921). *The Foundations of Science – Science and Hypothesis, The Value of Science, Science and Method* (G. B. Halsted, Trans.). The Science Press.

Popper, K. R. (1950). Indeterminism in Quantum Physics and in Classical Physics. Part I. *The British Journal for the Philosophy of Science*, *1*(2), 117-133.

Porter, S. L., Wadhams, G. H., & Armitage, J. P. (2011). Signal processing in complex chemotaxis pathways. *Nature Reviews Microbiology*, *9*(3), 153-165. https://doi.org/10.1038/nrmicro2505

Pospisil, D. A., Aragon, M. J., Dorkenwald, S., Matsliah, A., Sterling, A. R., Schlegel, P., Yu, S.-c., McKellar, C. E., Costa, M., Eichler, K., Jefferis, G. S. X. E., Murthy, M., & Pillow, J. W. (2024). From connectome to effectome: learning the causal interaction map of the fly brain. *bioRxiv preprint*. https://doi.org/10.1101/2023.10.31.564922

Potter, H. D., & Mitchell, K. J. (2022). Naturalising agent causation. *Entropy*, *24*(4), 472.

Potter, H. D., & Mitchell, K. J. (2025). Beyond Mechanism—Extending Our Concepts of Causation in Neuroscience. *European Journal of Neuroscience*, *61*(5), e70064. https://doi.org/10.1111/ejn.70064

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, *36*(1), 1-16. https://doi.org/10.1016/0010-0277(90)90051-K

Prigogine, I., & Stengers, I. (1984). *Order Out Of Chaos: Man's New Dialogue With Nature*. Bantam Books.

Pross, A. (2016). *What is Life? How Chemistry Becomes Biology* (Second Edition). Oxford University Press.

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, *14*(4), 180-190. https://doi.org/10.1016/j.tics.2010.01.008

Raja, V., & Anderson, M. L. (2021). Behavior considered as an enabling constraint. In *Neural mechanisms: New challenges in the philosophy of neuroscience.* (pp. 209-232). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-54092-0_10

Ramsey, W., Stich, S. P., & Garon, J. (1991). Connectionism, eliminativism, and the future of folk psychology. In W. Ramsey, S. P. Stich, & D. Rumelhart (Eds.), *Philosophical Perspectives* (Vol. 4, pp. 499-533). Lawrence Erlbaum.

Randi, F., Sharma, A. K., Dvali, S., & Leifer, A. M. (2023). Neural signal propagation atlas of Caenorhabditis elegans. *Nature*, *623*(7986), 406-414. https://doi.org/10.1038/s41586-023-06683-4

Ranti, C., Chatham, C. H., & Badre, D. (2015). Parallel temporal dynamics in hierarchical cognitive control. *Cognition*, *142*, 205-229. https://doi.org/10.1016/j.cognition.2015.05.003

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873-922. https://doi.org/10.1162/neco.2008.12-06-420

Raz, M. (2013). Alone Again: John Zubek and the Troubled History of Sensory Deprivation Research. *Journal of the History of the Behavioral Sciences*, *49*(4), 379-395. https://doi.org/10.1002/jhbs.21631

Redish, A. D. (2013). *The mind within the brain: How we make decisions and how those decisions go wrong*. Oxford University Press.

Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology*. Oxford University Press.

Reid, T. (1983). *Thomas Reid's Inquiry and essays* (R. E. Beanblossom & K. Lehrer, Eds. First Edition). Hackett Publishing.

Rich, P., Blokpoel, M., de Haan, R., & van Rooij, I. (2020). How Intractability Spans the Cognitive and Evolutionary Levels of Explanation. *Topics in Cognitive Science*, *12*(4), 1382-1402. https://doi.org/10.1111/tops.12506

Riskin, J. (2016). *The restless clock: a history of the centuries-long argument over what makes living things tick*. The University of Chicago Press.

Robinson, H. (2023). "Dualism". *The Stanford Encyclopedia of Philosophy (Spring 2023 Edition)*. URL = <https://plato.stanford.edu/archives/spr2023/entries/dualism/>

Robson, D. N., & Li, J. M. (2022). A dynamical systems view of neuroethology: Uncovering stateful computation in natural behaviors. *Current Opinion in Neurobiology*, *73*, 102517. https://doi.org/10.1016/j.conb.2022.01.002

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence [Original Research]. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/10.3389/fevo.2021.806283

Rosas, F. E., Geiger, B. C., Luppi, A. I., Seth, A. K., Polani, D., Gastpar, M., & Mediano, P. A. M. (2024). Software in the natural world: A computational approach to hierarchical emergence. *arXiv preprint*. https://doi.org/10.48550/arXiv.2402.09090

Rosas, F. E., Mediano, P. A. M., Jensen, H. J., Seth, A. K., Barrett, A. B., Carhart-Harris, R. L., & Bor, D. (2020). Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. *PLOS Computational Biology*, *16*(12), e1008289. https://doi.org/10.1371/journal.pcbi.1008289

Ross, L. N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal. *Synthese*, *202*(2), 56. https://doi.org/10.1007/s11229-023-04281-5

Ross, L. N., & Bassett, D. S. (2024). Causation in neuroscience: keeping mechanism meaningful. *Nature Reviews Neuroscience*, *25*(2), 81-90. https://doi.org/10.1038/s41583-023-00778-7

Ross, L. N., Jirsa, V., & McIntosh, Anthony R. (2025). The Possibility Space Concept in Neuroscience: Possibilities, Constraints, and Explanations. *European Journal of Neuroscience*, *61*(5), e70038. https://doi.org/10.1111/ejn.70038

Rosslenbroich, B., Kümmell, S., & Bembé, B. (2024). Agency as an Inherent Property of Living Organisms. *Biological Theory*, *19*(4), 224-236. https://doi.org/10.1007/s13752-024-00471-7

Rovelli, C. (2021). *Helgoland*. Allen Lane.

Rubel, A., Castro, C., & Pham, A. (2019). Agency Laundering and Information Technologies. *Ethical Theory and Moral Practice*, *22*(4), 1017-1041.

Rule, M. E., O'Leary, T., & Harvey, C. D. (2019). Causes and consequences of representational drift. *Current Opinion in Neurobiology*, *58*, 141-147. https://doi.org/10.1016/j.conb.2019.08.005

Runyan, J. D. (2024). Including or excluding free will. In M. Streit-Bianchi & V. Gorini (Eds.), *New Frontiers in Science in the Era of AI* (pp. 111-126). Springer Nature.

Rusakov, D. A., Savtchenko, L. P., & Latham, P. E. (2020). Noisy Synaptic Conductance: Bug or a Feature? *Trends in Neurosciences*, *43*(6), 363-372. https://doi.org/10.1016/j.tins.2020.03.009

Russell, E. S. (1924). *The Study of Living Things: Prolegomena to a Functional Biology*. Methuen & Company.

Ryle, G. (1949). *The concept of mind*. Barnes & Noble.

Sanborn, A. N., Zhu, J.-Q., Spicer, J., León-Villagrá, P., Castillo, L., Falbén, J. K., Li, Y.-X., Tee, A., & Chater, N. (2025). Noise in Cognition: Bug or Feature? *Perspectives on Psychological Science*, *20*(3), 572-589. https://doi.org/10.1177/17456916241258951

Sapolsky, R. M. (2023). *Determined: Life without free will*. Random House.

Sartorio, C. (2016). *Causation and Free Will*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198746799.001.0001

Sartorio, C. (2021). Indeterministic Compatibilism. In M. Hausmann & J. Noller (Eds.), *Free Will: Historical and Analytic Perspectives* (pp. 205-227). Springer International Publishing. https://doi.org/10.1007/978-3-030-61136-1_9

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, *55*, 103-111. https://doi.org/10.1016/j.conb.2019.02.002

Schlosser, M. E. (2008). Agent-causation and agential control. *Philosophical Explorations*, *11*(1), 3-21.

Schlosser, M. E. (2014). The luck argument against event-causal libertarianism: It is here to stay. *Philosophical Studies*. https://doi.org/10.1007/s11098-013-0102-1

Schlosser, M. E. (2019). "Agency". *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*. URL = <https://plato.stanford.edu/archives/win2019/entries/agency/>

Schlosshauer, M. (2005). Decoherence, the measurement problem, and interpretations of quantum mechanics. *Reviews of Modern Physics*, *76*(4), 1267-1305. https://doi.org/10.1103/RevModPhys.76.1267

Schlosshauer, M. (2014). The quantum-to-classical transition and decoherence. *arXiv preprint*. https://doi.org/10.48550/arXiv.1404.2635

Schlosshauer, M., Kofler, J., & Zeilinger, A. (2013). A snapshot of foundational attitudes toward quantum mechanics. *Studies in History and Philosophy of Science Part B:*

Studies in History and Philosophy of Modern Physics, *44*(3), 222-230. https://doi.org/10.1016/j.shpsb.2013.04.004

Schopenhauer, A. (1839/1999). *Prize essay on the freedom of the will* (E. F. J. Payne, Trans.; G. Zöller, Ed.). Cambridge University Press.

Schrödinger, E. (1944/2012). *What is Life? The Physical Aspect of the Living Cell: With Mind and Matter & Autobiographical Sketches*. Cambridge University Press.

Schurger, A., Sitt, J. D., & Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, *109*(42), E2904-E2913. https://doi.org/doi:10.1073/pnas.1210467109

Schwab, D. B., Casasa, S., & Moczek, A. P. (2017). Evidence of developmental niche construction in dung beetles: effects on growth, scaling and reproductive success. *Ecology letters*, *20*(11), 1353-1363. https://doi.org/10.1111/ele.12830

Searle, J. R. (2001). *Rationality in Action*. The MIT Press.

Sedley, D. (1983). Epicurus' Refutation of Determinism. *ΣΥΖΗΤΗΣΙΣ: Studi Sull' Epicureismo Greco e Romano Offerti a Marcello Gigante*, 11-51.

Seibt, J. (2016). "Process Philosophy". *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*. URL = <https://plato.stanford.edu/archives/win2016/entries/process-philosophy/>

Seitz, R. J., & Angel, H.-F. (2020). Belief formation—A driving force for brain evolution. *Brain and Cognition*, *140*, 105548. https://doi.org/10.1016/j.bandc.2020.105548

Semedo, J. D., Gokcen, E., Machens, C. K., Kohn, A., & Yu, B. M. (2020). Statistical methods for dissecting interactions between brain areas. *Current Opinion in Neurobiology*, *65*, 59-69. https://doi.org/10.1016/j.conb.2020.09.009

Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M., & Kohn, A. (2019). Cortical Areas Interact through a Communication Subspace. *Neuron*, *102*(1), 249-259.e244. https://doi.org/10.1016/j.neuron.2019.01.026

Seth, A. K. (2010). Measuring Autonomy and Emergence via Granger Causality. *Artificial Life*, *16*(2), 179-196. https://doi.org/10.1162/artl.2010.16.2.16204

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565-573. https://doi.org/10.1016/j.tics.2013.09.007

Shabo, S. (2011). Why free will remains a mystery. *Pacific Philosophical Quarterly*. https://doi.org/10.1111/j.1468-0114.2010.01388.x

Shabo, S. (2020). The Two-Stage Luck Objection. *Noûs*, *54*(1), 3-23. https://doi.org/10.1111/nous.12243

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791-806. https://doi.org/10.1016/j.neuron.2013.10.047

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475-1479. https://doi.org/10.1038/nn.2949

Shea, N. (2018). *Representation in Cognitive Science* (First Edition). Oxford University Press. https://doi.org/10.1093/oso/9780198812883.001.0001

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology*, *28*(7), 929-946.

Sherrington, C. S. (1910). Flexion Reflex of the Limb, Crossed Extension Reflex, and Reflex Stepping and Standing. *The Journal of Physiology*, *40*, 28-121. https://doi.org/10.1113/jphysiol.1910.sp001362

Shine, J. M. (2023). Neuromodulatory control of complex adaptive dynamics in the brain. *Interface Focus*, *13*(3), 20220079. https://doi.org/doi:10.1098/rsfs.2022.0079

Shine, J. M., Müller, E. J., Munn, B., Cabral, J., Moran, R. J., & Breakspear, M. (2021). Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nature Neuroscience*, *24*(6), 765-776. https://doi.org/10.1038/s41593-021-00824-6

Siemian, J. N., Arenivar, M. A., Sarsfield, S., Borja, C. B., Russell, C. N., & Aponte, Y. (2021). Lateral hypothalamic LEPR neurons drive appetitive but not consummatory behaviors. *Cell reports*, *36*(8), 109615. https://doi.org/10.1016/j.celrep.2021.109615

Silberstein, M. (2021). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences 2.0. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience* (pp. 363-393). Springer International Publishing. https://doi.org/10.1007/978-3-030-54092-0_16

Silberstein, M., & Chemero, A. (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science*, *80*(5), 958-970. https://doi.org/10.1086/674533

Silva, C., & McNaughton, N. (2019). Are periaqueductal gray and dorsal raphe the foundation of appetitive and aversive control? A comprehensive review. *Progress in Neurobiology*, *177*, 33-72. https://doi.org/10.1016/j.pneurobio.2019.02.001

Simon, H. A. (1955). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129-138.

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15-18). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20568-4_5

Sinnott-Armstrong, W. (2019). Contrastive mental causation. *Synthese*, *198*(3), 861-883. https://doi.org/10.1007/s11229-019-02506-0

Smart, J. J. C. (2003). Atheism and Theism. In *Atheism and Theism* (pp. 6-75). https://doi.org/10.1002/9780470756225.ch2

Smilansky, S. (2000). *Free Will and Illusion*. Oxford University Press. https://doi.org/10.1093/oso/9780198250180.001.0001

Smilansky, S. (2006). Review of Alfred R. Mele, *Free Will and Luck*. *Notre Dame Philosophical Reviews*, *2006*(11). https://ndpr.nd.edu/reviews/free-will-and-luck/

Smolin, L., & Verde, C. (2021). The quantum mechanics of the present. *arXiv preprint*. https://doi.org/10.48550/arXiv.2104.09945

Snell-Rood, E. C., & Ehlman, S. M. (2023). Developing the genotype-to-phenotype relationship in evolutionary theory: A primer of developmental features. *Evolution & Development*, *25*(6), 393-409. https://doi.org/10.1111/ede.12434

Solomon, P., Leiderman, P. H., Mendelson, J., & Wexler, D. (1957). Sensory deprivation: A review. *The American Journal of Psychiatry*, *114*, 357-363. https://doi.org/10.1176/ajp.114.4.357

Sourjik, V., & Wingreen, N. S. (2012). Responding to chemical gradients: bacterial chemotaxis. *Current Opinion in Cell Biology*, *24*(2), 262-268. https://doi.org/10.1016/j.ceb.2011.11.008

Spelke, E. S., Phillips, A., & Woodward, A. L. (1996). Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524021.003.0003

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*(2), 240-261. https://doi.org/10.1037/0033-2909.124.2.240

Steinemann, N. A., Stine, G. M., Trautmann, E. M., Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2024). Direct observation of the neural computations underlying a single decision. *eLife Sciences*. https://doi.org/10.7554/elife.90859.2

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., & Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, *576*(7786), 266-273. https://doi.org/10.1038/s41586-019-1787-x

Steward, H. (2009). Animal Agency. *Inquiry*, *52*(3), 217-231. https://doi.org/10.1080/00201740902917119

Steward, H. (2012). *A Metaphysics for Freedom*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199552054.001.0001

Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of Prediction. *Physical review letters*, *109*(12), 120604. https://doi.org/10.1103/PhysRevLett.109.120604

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, *75*(1), 5-24. https://doi.org/10.1007/BF00989879

Sultan, S. E., Moczek, A. P., & Walsh, D. (2022). Bridging the explanatory gaps: What can we learn from a biological agency perspective? *Bioessays*, *44*(1), 2100185. https://doi.org/10.1002/bies.202100185

Suzuki, M., Pennartz, C. M. A., & Aru, J. (2023). How deep is the brain? The shallow brain hypothesis. *Nature Reviews Neuroscience*, *24*(12), 778-791. https://doi.org/10.1038/s41583-023-00756-z

Szilágyi, A., Szabó, P., Santos, M., & Szathmáry, E. (2020). Phenotypes to remember: Evolutionary developmental memory capacity and robustness. *PLOS Computational Biology*, *16*(11), e1008425. https://doi.org/10.1371/journal.pcbi.1008425

Tagkopoulos, I., Liu, Y. C., & Tavazoie, S. (2008). Predictive behavior within microbial genetic networks. *Science*, *320*(5881), 1313-1317. https://doi.org/10.1126/science.1154456

Tavani, H. T. (2018). Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information*, *9*(4), 73.

Taylor, N. L., Whyte, C. J., Munn, B. R., Chang, C., Lizier, J. T., Leopold, D. A., Turchi, J. N., Zaborszky, L., Műller, E. J., & Shine, J. M. (2024). Causal evidence for cholinergic stabilization of attractor landscape dynamics. *Cell reports*, *43*(6), 114359. https://doi.org/10.1016/j.celrep.2024.114359

Terada, Y., & Toyoizumi, T. (2024). Chaotic neural dynamics facilitate probabilistic computations through sampling. *Proceedings of the National Academy of Sciences*, *121*(18), e2312992121. https://doi.org/doi:10.1073/pnas.2312992121

Thiele, A., & Bellgrove, M. A. (2018). Neuromodulation of Attention. *Neuron*, *97*(4), 769-785. https://doi.org/10.1016/j.neuron.2018.01.008

Thura, D., Cabana, J.-F., Feghaly, A., & Cisek, P. (2022). Integrated neural dynamics of sensorimotor decisions and actions. *PLOS Biology*, *20*(12), e3001861. https://doi.org/10.1371/journal.pbio.3001861

Timpe, K. (2016). Leeway vs. Sourcehood Conceptions of Free Will. In R. Kane (Ed.), *The Routledge Companion to Free Will* (pp. 213-224). Routledge.

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410-433. https://doi.org/10.1111/j.1439-0310.1963.tb01161.x

Toadvine, T. (2023). "Maurice Merleau-Ponty". *The Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*. URL = <https://plato.stanford.edu/archives/win2023/entries/merleau-ponty/>

Tognazzini, N. A. (2011). Understanding Source Incompatibilism. *Modern Schoolman*, *88*(1/2), 73-88.

Tononi, G. (2015). Integrated information theory. *Scholarpedia*, *10*. https://doi.org/doi:10.4249/scholarpedia.4164

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450-461. https://doi.org/10.1038/nrn.2016.44

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

Tseng, P., & Cheng, T. (2024). Causal prominence for neuroscience. *Nature Reviews Neuroscience*, *25*(8), 591-591. https://doi.org/10.1038/s41583-024-00838-6

Tsimring, L. S. (2014). Noise in biology. *Reports on Progress in Physics*, *77*(2), 026601. https://doi.org/10.1088/0034-4885/77/2/026601

Uller, T. (2022). Agency, goal-orientation and evolutionary explanations. *OSF preprint*. https://doi.org/10.31219/osf.io/49qrs

Uller, T., Feiner, N., Radersma, R., Jackson, I. S. C., & Rago, A. (2020). Developmental plasticity and evolutionary explanations. *Evolution & Development*, *22*(1-2), 47-55. https://doi.org/10.1111/ede.12314

Urai, A. E., Doiron, B., Leifer, A. M., & Churchland, A. K. (2022). Large-scale neural recordings call for new insights to link brain and behavior. *Nature Neuroscience*, *25*(1), 11-19. https://doi.org/10.1038/s41593-021-00980-9

Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and Machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0

van Bree, S., Levenstein, D., Krause, M. R., Voytek, B., & Gao, R. (2025). Processes and measurements: a framework for understanding neural oscillations in field potentials. *Trends in Cognitive Sciences*, *29*(5), 448-466. https://doi.org/10.1016/j.tics.2024.12.003

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford University Press.

van Inwagen, P. (1990). Logic and the free will problem. *Social Theory and Practice*, *16*(3), 277-290.

van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives*, *14*, 1-20.

van Inwagen, P. (2011). A Promising Argument. In R. Kane (Ed.), *The Oxford Handbook of Free Will*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399691.003.0024

van Inwagen, P. (2014). *Metaphysics* (Fourth Edition). Westview Press.

van Strien, M. (2014). On the origins and foundations of Laplacian determinism. *Studies in History and Philosophy of Science Part A*, *45*, 24-31. https://doi.org/10.1016/j.shpsa.2013.12.003

van Strien, M. (2021). Was physics ever deterministic? The historical basis of determinism and the image of classical physics. *The European Physical Journal H*, *46*(1), 8. https://doi.org/10.1140/epjh/s13129-021-00012-x

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive science and human experience*. The MIT Press.

Velleman, J. D. (1992). What Happens When Someone Acts? *Mind*, *101*(403), 461-481. https://doi.org/10.1093/mind/101.403.461

Vervaeke, J., Lillicrap, T. P., & Richards, B. A. (2012). Relevance Realization and the Emerging Framework in Cognitive Science. *Journal of Logic and Computation*, *22*(1), 79-99. https://doi.org/10.1093/logcom/exp067

Virenque, L., & Mossio, M. (2023). What is Agency? A View from Autonomy Theory. *Biological Theory*, 1-5.

Vogt, G. (2015). Stochastic developmental variation, an epigenetic source of phenotypic diversity with far-reaching biological consequences. *Journal of Biosciences*, *40*(1), 159-204. https://doi.org/10.1007/s12038-015-9506-8

von Bertalanffy, L. (1969). *General System Theory: Foundations, Development, Applications*. George Braziller.

Waddington, C. H. (1957). *The Strategy Of The Genes*. George Allen & Unwin.

Wagner, A. (2013). *Robustness and Evolvability in Living Systems*. Princeton University Press.

Walsh, D. M. (2006). Organisms as natural purposes: The contemporary evolutionary perspective. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *37*(4), 771-791. https://doi.org/10.1016/j.shpsc.2006.09.009

Walsh, D. M. (2015). *Organisms, Agency, and Evolution*. Cambridge University Press.

Walsh, D. M. (2018). Objectcy and Agency: Towards a Methodological Vitalism. In D. J. Nicholson & J. Dupré (Eds.), *Everything Flows: Towards a Processual Philosophy of*

Biology (pp. 167-186). Oxford University Press. https://doi.org/10.1093/oso/9780198779636.003.0008

Walsh, D. M., & Rupik, G. (2023). The agential perspective: Countermapping the modern synthesis. *Evolution & Development*, *25*(6), 335-352. https://doi.org/10.1111/ede.12448

Watson, R. A., & Szathmáry, E. (2016). How Can Evolution Learn? *Trends in ecology & evolution*, *31*(2), 147-157. https://doi.org/10.1016/j.tree.2015.11.009

Wegner, D. M. (2002). *The Illusion of Conscious Will*. The MIT Press. https://doi.org/10.7551/mitpress/3650.001.0001

Weinberg, R., Gould, D., & Jackson, A. (1979). Expectations and Performance: An Empirical Test of Bandura's Self-efficacy Theory. *Journal of Sport Psychology*, *1*(4), 320-331. https://doi.org/10.1123/jsp.1.4.320

West-Eberhard, M. J. (2003). *Developmental Plasticity and Evolution*. Oxford University Press. https://doi.org/10.1093/oso/9780195122343.001.0001

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Harvard University Press.

Winning, J., & Bechtel, W. (2018). Rethinking Causality in Biological and Neural Mechanisms: Constraints and Control. *Minds and Machines*, *28*(2).

Wittgenstein, L. (1953). *Philosophical Investigations*. Wiley-Blackwell.

Wolf, S. (1990). *Freedom Within Reason*. Oxford University Press.

Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*(2), 273-286. https://doi.org/10.1093/biostatistics/kxx069

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. https://doi.org/10.1093/0195155270.001.0001

Yoshihara, M., & Yoshihara, M. (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*(2), 53-64. https://doi.org/10.1080/01677063.2018.1468443

Yu, Z., Verstynen, T., & Rubin, J. E. (2025). How the dynamic interplay of cortico-basal ganglia-thalamic pathways shapes the time course of deliberation and commitment. *bioRxiv preprint.* https://doi.org/10.1101/2025.03.17.643668

Ziegler, J. F., & Lanford, W. A. (1979). Effect of Cosmic Rays on Computer Memories. *Science*, *206*(4420), 776-788. https://doi.org/doi:10.1126/science.206.4420.776

Zurek, W. H. (1991). Decoherence and the Transition from Quantum to Classical. *Physics Today*, *44*(10), 36-44. https://doi.org/10.1063/1.881293

Zutshi, I., Apostolelli, A., Yang, W., Zheng, Z. S., Dohi, T., Balzani, E., Williams, A. H., Savin, C., & Buzsáki, G. (2024). Hippocampal neuronal activity is aligned with action plans. *bioRxiv preprint*. https://doi.org/10.1101/2024.09.05.611533