The self-evidencing agent

Jakob Hohwy

Forthcoming MIT Press

[This is a preprint of the Introductory Chapter 1 to this book]

Jakob Hohwy
Monash Centre for Consciousness and Contemplative Studies
Monash University

Contents

Detailed Contents

Chapter 1. Introduction

Self-evidencing occurs when a model of the world exists because it is used to explain what is observed, such that the explained observations become the evidence for the model. This book argues that there is good reason to cast the existence of biological agents, such as humans, in terms of self-evidencing – that is, we are self-evidencing agents.

If we begin from self-evidencing, then we can use it as a first-principles method to understand what we are, how we experience the world and ourselves, and how we engage in diverse ways with the world. Self-evidencing can help situate attention, action, and perception. It also furnishes compelling approaches to decision-making and rationality, the self, and to preferences, volition, and values. It can build bridges to consciousness, and even on to free will, and to meaning and wisdom.

I want in this book to use self-evidencing to make sense of what I see as the whole, diverse spectrum of everyday experience, and our attempts to live well. We are embodied creatures, we are conscious, make decisions and enjoy some degree of autonomy. We sometimes connect strongly and fluidly to ourselves, to others and the world around us. Other times, our bodies betray us, we feel disconnected from who we are ourselves or from others and the world, and we feel out of control or controlled by circumstances. We can often perceive the world accurately, but what we see is also filtered through what we want and believe. We can be fully connected throughout the flow of an absorbing task, caught up in the moment, but can also be profoundly detached from the external world, lost in thought, mindwandering or dreaming. Though our individual human experiences have many features in common with other people's experiences, there are extraordinary differences and diversity in every person's lifelong trajectory through everyday experience, in our diverse sense of self and body, in the kinds of things we do to get what we each believe we want, and so on. Life is full of ever vacillating tensions and conciliations between the internal and the external, between self and other, between connection and detachment, and between body and mind, the conscious and the non-conscious, and freedom and determination. I am suspicious of philosophies that emphasise either the disconnected and internal, or the connected and embodied side of everyday experience. A good account of everyday experience should not paper over the full and diverse spectrum of our everyday experience.

Self-evidencing, as I shall develop it, allows the full gamut of everyday experience – warts and all. The simplicity of one process, self-evidencing, allows this lived cacophony to emerge from our causal connectedness to the world, but also puts the burden on the agent's ability to actively maintain their internal model of themselves and the world around them. The grounding perspective that self-evidencing offers is one of inescapable causal embedding but with profound epistemic solitude. Though the mind is highly active, the activity happens only within the system itself, blanketed behind our sensory organs, and having to do everything through a kind of model-constrained trial-and-error process, using variations in the mind's model of the world to elicit feedback as best we can from parts of our bodies and our environment. Given self-evidencing, this solitude is the inescapable reality of existence. This can seem a dire or pessimistic philosophy, providing a disembodied and scepticism-prone account of the mind. I think it is in fact an honest picture of the way we each manage to go through life. However, this view is at the same time also, more subtly and realistically, suffused with a modest optimism, enshrined in a belief that the world does afford our self-evidencing, and existence.

From self-evidencing flows several insights concerning things that matter to us. Self-evidencing demonstrates the tractability of our pursuit of truth and reason, in spite of the intractability of exact computational inference. It allows vast individual differences in everyday experience, and opens the door to our capacity for making wiser decisions and creating meaning in life, in spite of our diversity and occasional propensity for irrationality. It helps explain why we have ambivalent attitudes to risk, how we can explore and shape our own preferences and even our selves, how emotion and salience connect to who we are, and what it takes to bring value and meaning into our conscious lives.

In the last decade or so, the predictive processing framework has arisen for cognitive science, theoretical neurobiology, machine learning, and philosophy of mind. Predictive processing proposes that the brain engages in a kind of hypothesis testing, via precision-weighted predictive coding, which leads to perception. This was a dénouement of ideas going back to al Haytham, Kant, and Helmholtz presenting perception as a process of judgement or unconscious inference. Predictive processing has turned into a vibrant research paradigm, with many developments across computational theory, philosophy, experimental psychology and psychophysics, neuroimaging, and into many other domains as diverse as physiotherapy and aesthetics. One of the most important developments in the predictive processing framework is the focus on active inference, which is now seen as more central than standard predictive coding approaches. In active inference, action policies for planning and decision-making are inferred based on prior beliefs about which policy will best reduce uncertainty about future outcomes. There is a wealth of exciting work appearing on active inference, and it will play a pivotal role in this book.

Already back at the inception of the predictive processing framework it was clear that there were more ambitious principles at play, sitting behind the idea of the brain as an organ for prediction error minimisation. This was in the shape of the free energy principle, articulated and developed by Karl Friston. I then suggested ‘self-evidencing’, a concept borrowed from philosophy of science, as useful for speaking about these underlying principles. This book is provides a full-fledged unfolding of self-evidencing. The book explores philosophically what self-evidencing might tell us about the mind, going beyond the kinds of questions and topics mostly discussed for predictive processing and delving more broadly into everyday experience and things that, I think, matter greatly for our understanding of ourselves and how we try to live well.

Prediction error minimisation is intimately connected to self-evidencing, but predictive processing always conveyed a fairly idealised, uniform process of gradient descent on an error landscape, simply knocking down prediction error whenever it occurs, based on a cortical hierarchy that recursively repeats the same predictive coding mechanism at each level. In contrast, self-evidencing is a notion that affords a more nuanced and open-ended set of processes, where agents can engage in diverse ways as they maintain their internal model of their world and themselves in different contexts and through varying levels of uncertainty. This makes a new, broader set of tools available for approaching a wide range of topics, while still operating within the ontologically austere landscape of self-evidencing, provided by using the free energy principle as a method for understanding certain self-organising systems, namely ourselves. In this way, self-evidencing is an especially well-suited concept for advancing from the original notions of predictive processing, straddling technical and philosophical domains. Technically, to self-evidence is simply to maximise the marginal likelihood or Bayesian model evidence, through free energy minimisation. Philosophically, it opens for new epistemic and cognitive perspectives of relevance to many debates.[1]

Methodologically speaking, one can take what is sometimes (and non-pejoratively) labelled the 'low-road' approach to predictive processing. Here, different phenomena are considered one after another and we notice that often predictive processes such as predictive coding appear to be in play, either as perceptual inference or as active inference. On the winding 'low-road', the more phenomena we can explain with predictive processing, the more it seems an interesting framework. This is a fruitful and valuable way of doing science. It is consistent with treating predictive processing as an empirical hypothesis, which could for example express that certain organisms have evolved to use predictive processes to solve certain kinds of problems and thereby increase fitness.

A more theoretically ambitious framing is in terms of self-evidencing, grounded on the free energy principle. Here, predictive processing and active inference is motivated directly from a fundamental understanding of the existence of self-organising systems. It is this framing – or as it is sometimes called, 'the high-road' – that is in play in this book. On the less winding but more exhaustingly uphill high-road, self-evidencing is developed *a priori*, on the basis of conceptual analysis and mathematics. Here, the free energy principle is understood as similar to principles in physics, such as the principle of least action. It does not convey an empirical law of nature, which can be confirmed to various degrees through observation. The principle may or may not apply to certain systems when they are described in certain ways. In this sense, self-evidencing is a method or system for understanding existing things, such as humans. Empirical confirmation comes in when principles are applied to furnish process theories that describe particular systems, such as the anatomy and structure of brains and bodies of humans, or of other animals, or machines. Because self-evidencing is *a priori*, it is intended as a way to describe any self-organising thing that exists, and therefore, it would not make sense to say, for example, that self-evidencing is something that some organisms and not others have evolved in order to maximise fitness.[2]

It would be nice if there turns out to be copious empirical evidence for the process theories applying to human anatomy and cognition, which conform to self-evidencing and the free energy principle. I think the signs here are promising, though the empirical jury is still out. Though I will occasionally appeal to some of the empirical evidence acquired through process theories, the main thrust of the book will be the theoretical case built at the level of principles. I am mainly – and somewhat unapologetically – going by the conceptual, high-road case for self-evidencing. The motivation for this is two-fold: I think the *a priori* arguments for conceiving existence in terms of self-evidencing are compelling, and I think that comprehensively unfolding self-evidencing provides an important and worthwhile method for making sense of the diversity of everyday experience, while elucidating fundamental aspects of existence, ranging from rationality to value to consciousness and beyond.

This plan for the book means that I will discuss some classic philosophical topics, such as rationality, volition and free will, self, value, consciousness, wisdom, and meaning in life. To make this agenda tractable, I will make some simplifying moves for each topic. For example, I opt for compatibilism about free will, I allow that realism for the self is just for the self to be as real as other humdrum things, I am content with a naturalistic account of consciousness rather than a full-scale assault on the metaphysical mind-body problem, and I consider wisdom and meaning through their recent operationalisations in psychology, rather than their full unfolding through millennia of philosophy. When the explanatory targets are set up in these ways, self-evidencing can be deployed fruitfully, pointing to resolutions of long-standing debates, or helping to change our conceptions of what those debates are about. I hope building such bridges between self-evidencing and some of these long-standing debates can foster, and perhaps re-orient and unify, conversations within philosophy and across interdisciplinary domains.

There is plenty criticism of predictive processing, some philosophical or theoretical, and some empirical. Criticism is important and welcome for a thriving research program like predictive processing. I will not engage systematically with all the criticism here, which is a task I believe is better suited for more technical journal articles. But throughout the book, I will reference points of controversy and debate, suggesting ways of responding. Unsurprisingly, I am not swayed by the criticisms. When self-evidencing is seen in its full high-road scope as a first-principles method for philosophical investigation, it has ample resources to prevent the various criticisms from getting traction.[3]

Self-evidencing is in a peculiar dialectical situation because it is often touted quite imperiously by me and several others as the one principle to which all theories must conform. But that ambition is in fact consistent with a welcoming, collaborative dialectic, encompassing insights from other theories. Many theories from numerous scientific fields can be seen to conform to self-evidencing, without this conformity detracting from their empirical status or their theoretical insights. This opens the opportunity for synthesis and unification, where those theories extend our conception of self-evidencing, and self-evidencing can help transform and unify debates without rusticating all other theories to the trash heap.[4]

Self-evidencing then emerges as independently informative, but also as consistent with some existing approaches, and liberal with respect to various explanatory projects that researchers from various disciplines may embark on. That is, self-evidencing as a method wants to have its cake and eat it too – wanting to be both exclusive and inclusive. Self-evidencing accomplishes this by unification, demonstrating that scientific approaches to various problems (such as rationality, attention, self, volition, and so on) need not appeal to fundamentally distinct and potentially inconsistent theoretical constructs, somehow cohabiting within our cognitive systems. Importantly, self-evidencing under the free energy principle is a valuable method for describing existing biological agents, such as humans, because a self-evidencing agent can be understood as self-supervised – that is, able to perceive, attend, decide and act under their own steam and without relying on labelled training data or other external supervision. To me, this is significant because it means that any approach that conforms with self-evidencing has a chance of latching on to our fundamental existential condition, rather than presupposing that the computational challenges agents confront have already been solved by processes external to the agent. In contrast, approaches that cannot be brought into conformity with self-evidencing can never be truly self-standing, as they will always ultimately depend on more fundamental guarantees that the agents in question can be self-supervised. The consequence is that whereas self-evidencing is quite ecumenical, it does pose a challenge to the comprehensiveness and uniqueness of approaches that do not conform to it.

The book pursues self-evidencing into domains that touch on biology, physics, and neuroscience, and some of the key moves in the argument, which I seek to explain and frame philosophically, are recognisable from other fields, such as information theory, machine learning, dynamical systems theory, theoretical biology, and statistical physics. I also bring self-evidencing into debates about ethics and morality, and contemplative studies. The book primarily reflects my outlook as a cognitive philosopher, and I am exploring self-evidencing as a method for philosophical inquiry. It takes the story of self-evidencing to a certain conceptual stage, which will hopefully be a useful point of contact for experts in other fields such as biology or contemplative studies. To situate self-evidencing in the overall discourse relating to predictive processing, I reference numerous published studies and discussions, across multiple debates and disciplines. The literature on predictive processing is a burgeoning and I cannot capture it all, but I hope readers will be able to navigate from my reference points into the wider debates. To avoid cluttering the main text, references and more technical remarks about various topics and debates are collected in endnotes.[5]

Overall, the book is written as a first-principles argument for self-evidencing, and as advocating for self-evidencing as a fruitful method for approaching a broad, interdisciplinary array of debates about what we as self-supervised human agents experience and do. This project is in keeping with the excitement I and many others around the world have about predictive processing, active inference, and the free energy principle. That excitement is rapidly translating into amazing theoretical and computational advances, together with an increasing body of impressive empirical work on the process theories. Though the explanatory aims are ambitious for the book, I conceive the case I build here with relative humility, that is, the case is only as strong as each step of the *a priori* argument. I see the arguments developed in the book as providing both a philosophical backdrop for much of the ongoing theoretical and empirical research, and as pointing forward, charting a conceptual map for self-evidencing and its connection to debates that are central to our self-understanding.

• • •

Here is the plan of the book. Chapter 2 will bring out the core argument that self-evidencing is key to an analysis of existence. The chapter seeks to convey the peculiar status of the idea of self-evidencing for theorising about existence and mind, given its connection to the free energy principle. The first section provides a toolbox that will be useful for the entire book, with concise descriptions and examples of the notion of existence of things, and of self-evidencing and its key applications. A key insight in this chapter is that we only get an adequate method for investigating agents like us if it can be used to capture how our cognitive and perceptual processes are self-supervised (that is, the underlying process is tractable by the agent themselves, without the help of externally provided guidance, such as labelled training data). The free energy principle builds on variational methods for approximate inference, which I heuristically conceive as a kind of model-constrained trial-and-error process. The beauty of self-evidencing, with the free-energy principle, is that the agent's self-supervised trial-and-error processes are assessed against a simple, tractable objective function, namely the free energy (where free energy can be thought of as the fit between the observations expected under an assumed model of latent causes in the world and the observations they make, where the objective function will also be made to capture future observations and actions). The cost of that simplicity is that all the onus is on the agent's work to maintain their internal model of themselves and the world. That is where the complexity lies and where diversity among individuals emerge.

Chapter 3 takes the core conception of self-evidencing, understood in the light of the free energy principle, and asks what a self-evidencing organism might look like. The starting point here is how self-evidencing is about maximising model evidence, and how this evokes an idea from philosophy of science about inference to the best explanation. The chapter builds a case that inference to the best explanation leads to just the processes entailed by the free energy principle – balancing complexity and accuracy, and risk and ambiguity. This helps us see what is distinctive about self-evidencing. The chapter further develops a key theme in the free energy principle, namely that self-evidencing offers a synthesis or conciliation of descriptive and normative perspectives on cognitive processing. Finally, the chapter puts forward-looking processing at the centre stage, highlighting allostasis and introducing active inference.

Chapter 4 revisits themes that have been central to predictive processing, namely perception, attention, and the role of action in perception. In my previous book, *The Predictive Mind*, I told a mechanistic story beginning with predictive coding for perception, then casting attention as precision-weighted predictive coding, and adding active inference to account for ways in which action changes attention and perception. However, in light of self-evidencing, these discussions are re-organised and reconceived. Precision is now the basic mechanistic notion, understood in the setting of active inference as precision control. From precision control, attention, action and perception flows. The general outlook is still that existing organisms must minimise uncertainty but now this is forward-looking, focusing on expected uncertainty, before the inferences then unfold in real-time. This re-ordering has deep implications for understanding what we do and what mental life is. It presents self-evidencing agents as epistemically solitary, yet causally embedded in the world.

Chapters 2, 3 and 4 thus set out the conception of the self-evidencing agent. The following chapters then explore self-evidencing through several themes and areas of debate, showing how in each case self-evidencing can unify and illuminate our understanding.

"Self-evidencing" might appear perniciously circular. How can such a seemingly circular construct lead to accurate beliefs, perceptions, and representations? How can it be conducive to getting us what we desire, if it is an epistemically flawed, circular construct? Chapter 5 brings out how self-evidencing facilitates accurate representation, noting that this kind of accuracy is consistent with a pragmatic, self-serving aspect of self-evidencing, forming an attractive package. This leads to useful insights on other kinds of apparently false inference as well as more general sceptical challenges to knowledge. Existence as self-evidencing carves out a comfortable and productive middle position between never being wrong and always being wrong. This chapter also considers the accuracy of the self-evidencing framework itself, considering its *a priori* status and role in an empirical research program.

Following on, Chapters 6, 7 and 8 each use the fundamental notion of self-evidencing to develop novel perspectives on key concepts of rationality, self, and volition and value.

Chapter 6 considers decision-making and rationality from the self-evidencing perspective. Self-evidencing provides, I will argue, an attractive new idea of rationality, modelled on inference to the best explanation. This is inference to the best decision, which provides a coherent and accommodating blend of descriptive accounts of tractable rationality and normative accounts of intractable, ideal rationality. Here, decision-making is based on a particular kind of model-constrained trial-and-error process, rather than on predetermined algorithms or on bare heuristics. I liken rationality to a kind of abductive hill-climbing, allowing much diversity among individuals.

Chapter 7 looks at what happens to the self in self-evidencing, distinguishing between the self-model and the broader internal model of the world. There is an important point to convey first for this chapter, namely that the 'self' in 'self-evidencing' refers in the first instance to the circular process of a model explaining some observation that becomes evidence for the model it*self*. The earlier chapters argued that the overall internal model can be equated with the agent, since if the agent exists it must be a self-evidencing model. Chapter 7 argues that what we call 'self' is a subset of the internal nodes of this broad model, representing those causes of our observations that originate within us. This makes the self as real as any other thing we deem real. The chapter brings out how we can close the action-perception loop to provide evidence for the specifics of the model of our own self, consolidate and clarify it, and perhaps change it – shaping ourselves. At this point, a picture of specifically human self-evidencing begins to emerge on which we are, as I shall label it, agents of volatility. Such agents occupy and exploit an eco-niche characterised statistically by volatility, or relatively abrupt state transitions. The notion of agents of volatility then informs parts of the following chapters.

Self-evidencing presents agents as wholly enmeshed in the causal order of the world around them, and operating through a mechanism of internal causal precision control. In Chapter 8, I ask what happens to volition, emotion and value on such a thoroughgoing causal picture. Much work on decision-making assumes that we act for value or utility, occasionally with some added room for acting for epistemic value. I discuss how self-evidencing entails that resolving and reducing uncertainty is the fundamental outcome of action, from which utility and epistemic value naturally emerge, making sense of many of our patterns of behaviour. I discuss how the sense of volition, consistent with compatibilist conceptions of free will, plays into this picture, in particular for agents of volatility such as humans. I then discuss issues of emotions and sensation, building on several existing discussions. The chapter finishes by diving into areas of ethics and value. The question there is if self-evidencing, with its unwavering focus on the self, might be inherently egotistical? I argue that value and ethics cannot arise inherently from self-evidencing, but that self-evidencing nevertheless has a place for them, emphasising the role of democratic discourse. I also speculate that some of the self-evidencing tools can help navigate an interesting route in some contemporary debates about utilitarianism.

In Chapters 9 and 10, several of the pieces from previous chapters are brought together to elucidate substantive debates on consciousness, and on meaning and wisdom.

In Chapter 9, I discuss how the active inference element of self-evidencing can be brought to bear in consciousness science. This chapter both prepares the ground such that self-evidencing as a method can apply meaningfully to consciousness science, and it lays out how the various elements from previous chapters together provide what I think is a fruitful account of consciousness. Thus, I first advocate for a particular methodological approach to the thorny issue of consciousness, which can both reveal underlying mechanisms for conscious phenomenology and let us glimpse why consciousness has its particular subjective feel; however, I make clear that such an account is not a solution to the hard, metaphysical problem of consciousness (and this is a good thing). I then build a case that active inference is necessary for changes in conscious content, providing an illuminating account of the underlying mechanisms of consciousness. My case proceeds by showing how the counterfactual processing of active inference, in agents of volatility in particular, provides the detachment that marks out our conscious experience. I then outline how active inference can explain conscious content in everyday perception, and how this plays out as the policies selected in active inferences are enacted. I finish by highlighting elements of active inference that speak to subjectivity and unity of consciousness. Overall, self-evidencing begins to unify self, volition, and consciousness, within agents of volatility.

The austere, *a priori* analysis of existence underpinning self-evidencing may seem a long way from human well-being and flourishing. Chapter 10 therefore considers some quintessential philosophical questions about wisdom and meaning, filtered through contemporary psychological debates. Can self-evidencing be used to re-cast ancient aspirations to be not just clever or rational but also to be wise? Can it be used to explain the nebulous idea that we can experience, and perhaps enhance, meaning in life? I will argue that these otherwise somewhat exalted notions of meaning and wisdom can be harnessed by self-evidencing, and shown to be achievable in diverse ways for us all. The chapter ends with a more practical perspective, picking up on internal precision control as a cognitive mechanism for mindfulness.

The concluding remarks in Chapter 11 collects the threads, summarising why self-evidencing makes sense of mind and cognition, and the diversity of everyday experience. We are cast into existence and keep ourselves afloat by, calmly or frantically, adjusting the beliefs of our internal model. We do this secluded behind the veil of our senses, and can only rely on a trial-and-error process based on our own beliefs about the world and ourselves, and assessed only against overall uncertainty. We manage our way through, sometimes going well, other times accumulating more error than success. Though self-evidencing is inherently solitary, it optimistically allows knowledge, rationality, self, emotion, value and consciousness. Though the self-evidencing agent can fail miserably, they can also live well, continuously learn how to act more wisely and how to create more meaning in life.

Bruineberg, J., Dołęga, K., Dewhurst, J., and Baltieri, M. 2022. The Emperor's New Markov Blankets, *Behavioral and Brain Sciences*, 45: e183.

Clark, A. 2013. Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science, *Behavioral and Brain Sciences*, 36: 181-204.

———. 2016. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. New York: Oxford University Press.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory?, *Nature Reviews: Neuroscience*, 11: 127-38.

Friston, K. 2019. Beyond the Desert Landscape. In *Andy Clark and His Critics*, 174-90. Oxford University Press.

Friston, K., Da Costa, L., Sakthivadivel, D.A.R., Heins, C., Pavliotis, G.A., Ramstead, M., and Parr, T. 2023. Path Integrals, Particular Kinds, and Strange Things, *Physics of Life Reviews*, 47: 35-62.

Hohwy, J. 2013. *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. 2016. The Self-Evidencing Brain, *Noûs*, 50: 259-85.

———. 2020. New Directions in Predictive Processing, *Mind & Language*, 35: 209-23.

———. 2021. Self-Supervision, Normativity and the Free Energy Principle, *Synthese*, 199: 29-53.

Nave, K. 2025. *A Drive to Survive: The Free Energy Principle and the Meaning of Life*. Cambridge, MA.: MIT Press.

Parr, T., Pezzulo, G., and Friston, K.J. 2022. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. Cambridge, Mass.: MIT Press.

Raja, V., Valluri, D., Baggs, E., Chemero, A., and Anderson, M.L. 2021. The Markov Blanket Trick: On the Scope of the Free Energy Principle and Active Inference, *Physics of Life Reviews*, 39: 49-72.

Ransom, M., Fazelpour, S., and Mole, C. 2017. Attention in the Predictive Mind, *Consciousness and Cognition*, 47: 99-112.

Sun, Z., and Firestone, C. 2020. The Dark Room Problem, *Trends in Cognitive Sciences*, 24: 346-48.

Williams, D. 2022. Is the Brain an Organ for Free Energy Minimisation?, *Philosophical Studies*, 179: 1693-714.

Chapter 1

[1] My paper on self-evidencing appeared in *Noûs* in 2014, published as (Hohwy 2016). In this book there will be many references to the predictive processing literature, with a slant toward philosophical treatments. My own book (Hohwy 2013) and Andy Clark's influential paper and subsequent book (Clark 2016, 2013) can be useful starting points. Already by 2020, a large body of philosophically oriented work on predictive processing had emerged, with several hundred listed in the appendix to a review I conducted (2020) and only growing since. Karl Friston articulated the free energy principle, and is the main source, see for example Friston's extremely influential landmark paper (Friston 2010) and the many other works by Friston, many of which I reference in this book.

[2] Concerning principles specifically, the free energy principle is a variational principle of least action; the recent technical literature thus makes it clear that the free energy principle implies that there exists an action for every existing system; namely, the path integral of variational free energy (Friston et al. 2023). For discussion of the 'high-road' vs. 'low-road' approaches to predictive processing, see (Friston 2019); for discussion of principles and processes, and references to the philosophy of science debates about

the relation between principles and laws, see (Hohwy 2021), and for critical discussion of the distinction, see (Williams 2022). The metaphor of the high and low roads picks up on the idea that both will end up at the same destination, via different methodologies. Which methodology one prefers depends on pragmatic interests and scientific choices; the metaphor as it is employed in the literature on the free energy principle is not intended to say that one is better than the other.

[3] For many of the more philosophical critiques of predictive processing and the free energy principle, there are already several published responses, which, as I see it, efficiently engage with the criticisms. See for example the commentaries by Friston, Ramstead, Andrews, Parr, Kiefer and myself, Seth, Clark, Kirchhoff and several others published with the following critiques of predictive processing (Bruineberg et al. 2022; Raja et al. 2021; Sun and Firestone 2020; Ransom et al. 2017). There are several further published critical papers, many of which I will allude to along the way.

[4] There are several formal accounts of how other frameworks fit in with the free energy principle, see for example (Friston 2010; Parr et al. 2022).

[5] As a companion to this book, I recommend the excellent and authoritative *Active Inference* by Thomas Parr, Giovanni Pezzulo and Karl Friston (2022). For a comprehensive and detailed counterpart, I recommend Kathryn Nave's *A drive to survive* (2025), which takes a more embodied or enactivist approach to the free energy principle.