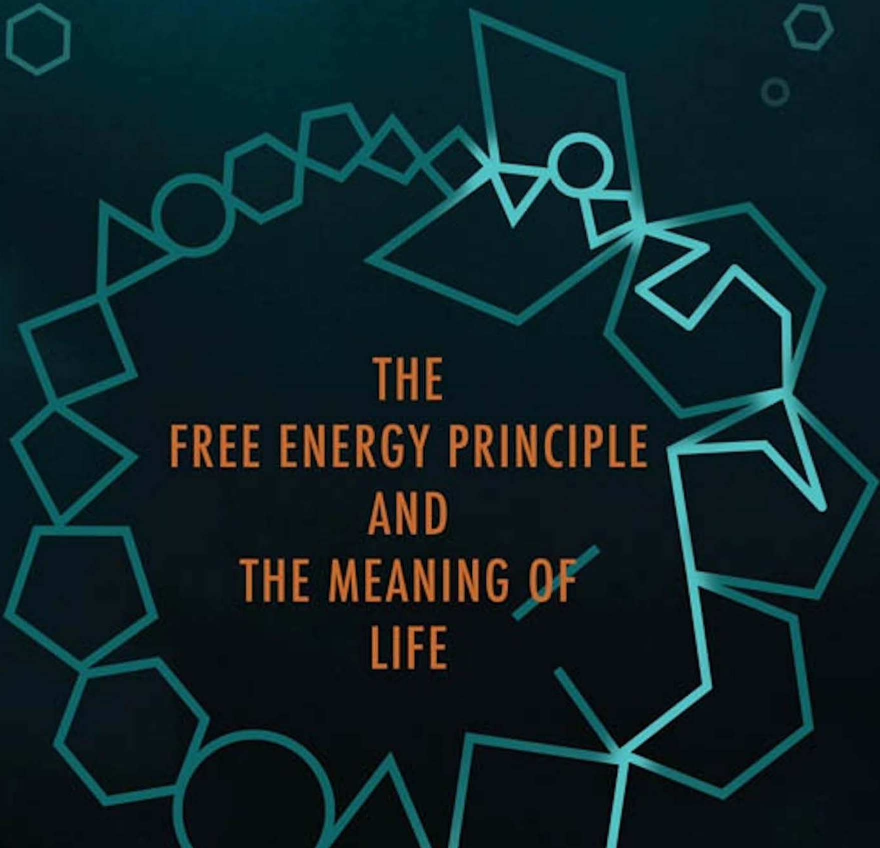
A DRIVE TO SURVIVE



THE FREE ENERGY PRINCIPLE AND THE MEANING OF LIFE

KATHRYN NAVE

A Drive to Survive

A Drive to Survive

The Free Energy Principle and the Meaning of Life

Kathryn Nave

The MIT Press
Cambridge, Massachusetts
London, England

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Nave, Kathryn, author.
Title: A drive to survive : the free energy principle and the meaning of life / Kathryn Nave.
Description: Cambridge : The MIT Press, [2025] | Includes bibliographical references and index.
Identifiers: LCCN 2024017278 (print) | LCCN 2024017279 (ebook) | ISBN 9780262551328 (paperback) | ISBN 9780262381666 (epub) | ISBN 9780262381673 (pdf)
Subjects: LCSH: Intentionalism. | Inference. | Agent (Philosophy) | Autonomy (Psychology) | Phenomenology.
Classification: LCC BF619.5 .N38 2025 (print) | LCC BF619.5 (ebook) | DDC 153—dc23/eng/20240722
LC record available at https://lccn.loc.gov/2024017278
LC ebook record available at https://lccn.loc.gov/2024017279

Contents

Acknowledgments

This book has been directly shaped by the following people: Dave Ward, Andy Clark, Alistair Isaac, Mazviita Chirimuuta, and Evan Thompson who read parts of this manuscript in its earlier stages and provided encouragement and useful feedback for its improvement. Particular thanks to Evan for helping me to get it published. I'd also like to thank Jaan Aru, whose lovely and unexpected email gave me the final push I needed to actually send out a book proposal. This manuscript is also much better thanks to three anonymous reviewers and to Giovanni Pezzulo—all of whom were simultaneously generous in their assessments and insightful in their suggestions for improvement. The Royal Institute of Philosophy, The Aristotelean Society, The Analysis Trust, and the Leverhulme Trust all funded various stages in the preparation of this book. Without their support, it would not exist at all.

I've been lucky enough to spend the past six years at Edinburgh University's school of philosophy, psychology, and language sciences, learning both from and with the following wonderful people: Lilith Lee, Matt Sims, Danaja Rutar, Jodie Russell, Emma Otterski, Shannon Prosckh, Jennny Zhang, Mark Sprevak, Frank Schumann, Mog Stapleton, and many others. The people who have influenced and inspired me through various conferences and summer schools are too numerous to list, but to the extent I've been able to talk coherently about Markov blankets, I thank Benjamin Jantzen's incredible summer school on philosophy and physical computing.

While the technicalities of the free energy principle have preoccupied me for only the past few years or so, I've been trying to figure out exactly what's going on since I was first puzzled by how my hand could seem like both an object in the world and a part of my self. The fact that I'm now paid to think about effectively the same question is thanks to the support of the following people: my parents, Kym West and Colin Nave, who never told me to stop wasting my time and do something more sensible; my partner, Max Wilkinson, who regularly does and who drags me away from the theoretical world to come play in the real one; Kate Webb, who is always up for exploring both at the same time; Amy Mallinson, who is more of a philosopher than she gives herself credit for; Mr Jackson, who was willing to continue teaching a philosophy class, even with only one student; Graham, Louise, Izzy, and Chris, who welcomed me into their family and made numerous valiant efforts to understand what I'm talking about; the climbing fam who don't care as long as I'm up for another trip to Font; and whatever genes I inherited from my grandad, who I know only through his collection of popular philosophy books. It would have been a thrill to put this on your shelf.

Introduction: Control Is Not the Goal

> Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals.
>
> —The European Commission's High-Level Expert Group on Artificial Intelligence (2018)

> I am less disturbed by a science that claims to be the equal of God than by a science that drains one of the most essential distinctions known to humanity since the moment it first came into existence of all meaning: the distinction between that which lives and that which does not; or, to speak more bluntly, between life and death.
>
> —Jean-Pierre Dupuy (2018)

0.1 Introduction

This book could not have been written without support from the inhabitant of a large basement that lies just beneath where the M4 motorway passes through Wokingham. Here lives the National Grid's Electricity Control Centre (ECC), the "brain" behind the UK's power network. By monitoring millions of sensors across the country, this partially automated system works to continuously maintain a level of 400 kilovolts (kv) throughout the grid's five thousand miles of powerlines. The maintenance of an uninterrupted flow of electricity to the computer at which I'm writing, alongside every other electrical system in the country, depends upon the ECC's success at mantaining that set point within a tolerance of just five percent.

If supply outstrips demand, the ECC releases pressure by lowering electricity prices. If demand increases, the control center rapidly spins up its reserves to compensate—for instance, by instructing Dinorwig power station to release the nine billion gallons of water stored near the summit of Elidir Fawr, converting this gravitational potential energy into enough electrical power to ensure that millions of households across the UK can simultaneously make their morning cup of coffee without plunging the nation into a blackout.

In recent years, this regulatory task has become more challenging, thanks to the replacement of traditional voltage support reserves, such as coal and gas, with the environmentally friendlier but inconstant alternatives of solar and wind. In order to budget with such unreliable forces, the grid has had to get "smarter" via the deployment of AI systems to forecast demand increases or weather-induced outages. These enable the ECC to automatically take predictive actions to maintain the delicate balance of the grid. There is no in-principle reason why the national grid's control center could not become a fully automated, anticipatory system.

The ECC is, in a weak sense, dependent on its own successful operation. If supply is not matched to demand, then on-site backup generators will sustain its regulatory system's operation throughout the consequent blackout for only so long before the system runs out of the power needed to continue operating. As the civil engineering professor Guy Walker (2013) describes, it is "more like an organism than a machine" (p. 252).

The British cybernetician William Ross Ashby would have endorsed the comparison. For Ashby, this homeostatic regulation of "essential variables" is all that there is to being a living system, or indeed, to being any sort of system at all. Survival means nothing more than stability amid perturbation, and the only difference between an organism learning to adapt to a new environment, the regulatory activity of the ECC, and the action of a pendulum returning to its equilibrium point is the complexity of the restabilization mechanisms involved. As he puts it:

> We have heard ad nauseam the dictum that a machine cannot select; the truth is just the opposite: every machine, as it goes to equilibrium, performs the corresponding act of selection. Now, equilibrium in simple systems is usually trivial and uninteresting; it is the pendulum hanging vertically; it is the watch with its mainspring run down; the cube resting flat on one face . . . What makes the change, from trivial to interesting, is simply the scale of the events. (Ashby, 1962, p. 70)

The prevailing currents of twentieth-century artificial intelligence were not kind to Ashby. It was far from obvious to his fellow cyberneticians how a machine like his "homeostat," which randomly flailed around when disrupted until it reobtained a stable state—a "sleeping machine," as Ashby's fellow cybernetician Grey Walter caricatured it—might one day "play chess with a subtlety and depth of strategy beyond that of the man who designed it" (Ashby, 1948).

The intelligence researchers of the late 1900s preferred the clean-shaven respectability of algorithmic symbol manipulation. Accordingly, much work in AI and cognitive science focused on characterizing the computational "means" of intelligence rather than the end. The problem is that, in purely formal terms, computation is trivial and any physical assemblage of moving parts, whether a pendulum or a difference engine, can be interpreted as executing an algorithmic operation. If we want to distinguish intelligent systems that are genuinely adding, subtracting, inferring, or integrating, then a syntactic account of computation cannot be prioritized over an account of the meaning of these formal operations, or the function they perform. Ends must come before means.

This century has proved more receptive to the idea of intelligence as control and a focus on the achievement of ends rather than an understanding of the means. Today, artificial intelligence is almost synonymous with deep learning where success is measured by the reliability with which the output of a many-layered network converges on a value that we've determined as desirable. The operations by which such networks achieve this are often impenetrable, and it is precisely this opacity, with the occasional unexpected behavior that results, that seems to motivate our perception of them as "intelligent." The European Commission's definition of artificial intelligence at the beginning of this introduction might well be reformulated without loss of meaning as "a system that does something we find useful in a manner we find hard to understand."

So, we speak of a complex regulatory system like the ECC as "autonomous" or "smart," we credit it with "trying" to achieve goals, and we get mad at it when it fails. For Ashby, such a system is thus "heaven-sent in this context, for it enables us to bridge the enormous conceptual gap from the simple and understandable, to the complex and interesting. Thus we can gain considerable insight into the so-called spontaneous generation of life by just seeing how a somewhat simpler version will appear in a computer" (Ashby, 1962, p. 271).

As applications of this control-first approach have begun to deliver upon the practical consequences of Ashby's proposal, with the creation of artificial intelligences that can not only play chess but write poems about it, so cognitive scientists, working on the ideas of "active inference" and "predictive processing," have begun to turn their attention to this theoretical task: the attempt to understand the emergence and operation of biological intelligences through the same lens. According to this predictivist turn, the difference between a simple homeostat and a human brain is simply a matter of the degree to which the latter facilitates a form of *predictive* control over multiple timescales. This ability to coordinate our actions in anticipation of *expected* disruptions is then proposed to underlie the kind of sophisticated capacities such as inference, imagination, strategization, and exploration, which distinguish natural intelligence.

Still, such work is less commonly inspired by Ashby and the cyberneticists than by Karl Friston's more recent free energy principle (FEP), initially introduced in 2005 as a "theory of cortical functioning" and since extended to explain everything from "cells to societies" in terms of a principle of homeostasis (Friston, 2005; Ramstead et al., 2021). In scope and content this is as Ashbyian as it gets. Unlike Ashby, however, who saw his program as the *elimination* of teleological interpretations from the scientific explanation of intelligent behavior (biological or otherwise), Friston and his various co-authors have explicitly advocated the description of these control hierarchies in purposive terms. By interpreting predicted states as "goals" that a hierarchical control architecture is "trying" to bring about, they claim that they can describe the sense in which such a system has *purposes* and the intentions to bring them about (Ramstead et al., 2018).

If we consider a system's predictive goal as being the stable state in which it is most likely to be found, then its return to this state when disrupted becomes an act of control via the minimization of prediction error. Yet, as Ashby emphasized, everything from a pendulum to a watch spring "rejects" unstable states to "select" a stable equilibrium. If intelligence, agency, and intentions reduce to nothing more than this form of predictive control, then they are either everywhere or nowhere at all.

We may take some small comfort in the reminder that our own brains and their prediction-error minimizing behavior are vastly more complex in their capacity for restabilization than a simple pendulum—but what of the ECC? Is my striving toward the goal of finishing this book no more significant than its increasingly complex anticipatory operations toward the "goal" of maintaining that 400 kv set point?

Arguably, much more rides on the ECC's success. But while it would certainly matter to us if the National Grid suffered a power loss, does the

ECC itself care? There is, I suggested, a sense in which the ECC depends on maintaining the grid's 400 kv supply. Its regulatory operations require electricity, and absent its own activity in ensuring they receive this, those operations would eventually cease. The ECC is not particularly novel in this respect. The planet's hydrological cycle constitutes a similar cycle of mutually dependent processes, yet it is no longer fashionable to attribute wrathfulness, or other agential attributes, to the weather.

While the ECC is not so unlike the hydrological cycle, as I will attempt to show in this book, it is quite fundamentally different from an organism. In a living cell, there is not only dependence of activity on activity but of existence on activity. The ECC's physical parts, its silicon transistors and copper wires, are intrinsically stable and will not automatically disintegrate if energy ceases to run through them. Reintroduce a supply of electricity, and it can continue operating in just the same manner as before. In contrast, the internal enzymes and surrounding membrane of a cell are inherently less stable than the cell itself and, as such, are dependent on the very metabolic activity that they enable for their ongoing repair and replacement. Deprived of the supply of matter and energy that fuels this activity, the very structure of the organism will irreparably disintegrate.

The ECC is a particularly complex stabilization mechanism, but a stabilization mechanism is all that it is. We call this stable state of 400 kv a “goal” simply because it is reliably achieved and desirable to us, and we call the ECC, but not the egg-timer, “intelligent” merely because we do not understand it.

Perhaps this is all there is to being an intelligent agent. Perhaps talk of goals and intentions are convenient heuristics to abstract away from the messy mechanical details of a system's operation. This instrumentalist view of intelligence and agency is the only one available within a mechanistic perspective, where the stability of parts is presumed and we are interested only in how they produce a particular behavior.

But the mechanistic does not exhaust every possible form of existence. The bacteria swarming all over my keyboard may not seem particularly impressive in comparison to the ECC's ability to probabilistically model the likelihood of a variety of future events and allocate resources in anticipation. Unlike the ECC, however, a bacterium can justifiably take credit for its own existence. Unlike the ECC, its structure is precarious and reciprocally dependent on the activity it produces. The bacterium cannot be cut off from energy flows without consequences for its physical integrity. Unlike the ECC, the bacterium does not merely respond to perturbations to an otherwise stable state: it is intrinsically dynamic as the inevitable degradation of its internal components releases energy to drive the activity that rebuilds them.

The organism, as the philosopher Hans Jonas put it, has a "needful freedom" in relation to matter—both independent of any particular material basis and dependent on a continual supply in order to continuously reproduce itself. Energy flows through the fixed structure of the ECC, but in the organism everything flows.

The bacterial colony of my keyboard must reliably maintain certain metabolic processes as an existential imperative. If they do not achieve the necessary flows of matter and energy, they will not only cease to be active—they will cease to exist. To say the bacterium needs to constantly rebuild itself in this way is not an anthropocentric projection. It has nothing to do with us at all.

So, to be a realist about purposes, intentions, or goals is not vitalistic or unscientific. It is to attend to the natural and intrinsic features of biological existence that are erased by a purely mechanistic conception of the universe. Taking the function of a system to be instrumental to our purposes, as the mechanist often does, was never a particularly satisfying solution in the first place. At some point, this instrumentalist will have to explain what's so special about us, in virtue of which we can have needs and purposes for other things to be relative to.

If this "bioenactive" view of what it means to be an intentional agent supplies the ends of intelligent behavior, the question remains as to how complex and creative intellectual capacities could arise from such a simple imperative of survival. Hierarchical predictive control, I argue, serves this biological purpose well enough in helping us anticipate and avoid threats that might disrupt our ongoing metabolic self-production. Insofar as such predictive architectures have also been advanced as a potential explanation for how we reason about everything from future actions to other peoples' mental states, so this framework may serve to show how the basic biological goal of survival could—through evolution, learning, and social scaffolding—lead to the catching of baseballs, the dancing of tangos, the arranging of flowers, or the scaling of mountains.

In making the processes of cognition less opaque, perhaps such models will also make them seem less intelligent. But if the intelligence we ascribe to another system is only a reflection of our own ignorance, I'm not sure it was all that worth caring about in the first place. Whether a hierarchical predictive model is autonomous, whether it is an agent with needs and investment in its own activity, depends not on our perspective but on whether it is a vital constituent of an intrinsically unstable system that is only temporarily stabilized by its own operations. Control is the means, not the goal.

You can grant the ECC access to every power line across the globe and install the most sophisticated predictive algorithms and the largest of large language models. You can hook it up to a speech synthesizer and instruct it to sing of its feelings for electricity, install Windows 95 and use it to play *Doom*, or tear out its silicon chips to make earrings and sell them on Etsy. Either way, the ECC itself will not care. No matter how complex it becomes and the variety of perturbations that it is able to achieve a stable state in anticipation of, insofar as that stable state is a "goal," it will only ever be ours.

Unlike the ECC, the bacterium needs to seek out continual flows of matter and energy from its environment to continue to exist. You probably don't care if it succeeds. The bacterium likely lacks the counterfactual flexibility or recursive self-modeling required to care much either. But unlike the ECC, it has something that it could, at least in principle, learn to care about.

If intelligence is about pursuing goals, then, I argue, it has to start here.

0.2 Overview

In the first chapter, I introduce the enactive approach and contextualize it as a naturalistic continuation of the phenomenological tradition, initiated by the philosopher Edmund Husserl in the early twentieth century. Set in this context, the enactive approach's primary goal can be understood as the attempt to supply an alternative *teleological* conception of intentionality, in terms of the striving of a system toward some nonreconstructive goal or norm. To supply this, I advocate what I'll refer to as "*bio*enactivism," which aims to locate this immanent teleology in the autonomy distinct to living systems, whose precarious existence is dependent on their own activity in working toward their continuous self-production.

While this may give us a foundation for normative evaluations, as Di Paolo (2005) has argued, it is only the all-or-nothing normativity of continued existence. To attribute a graded normativity to an agent's interactions with the world, we need to introduce what he terms "adaptivity": the capacity of a system to regulate these interactions so as to move toward, or away from, states according to whether they threaten, or support, its autonomous organization. While Di Paolo et al. (2017) have progressed the enactive account by describing how this adaptivity might be "scaled up" through the process of sensorimotor equilibration, accounts of how this is implemented are still needed.

This, as I propose in chapter 2, is where predictive processing (PP) and its analysis of the brain as a system for hierarchical prediction error minimization, can come in. When presented as a model of how we can extract patterns from sensory input over multiple timescales to develop predictive models of the relationships between inputs and outputs, PP looks to do away with what Hurley (1998) termed the "classical sandwich" view of cognition, as an independent operation that goes on in between the separate processes of perception and action (Vázquez, 2020; Nave et al., 2020; Harvey, 2018; Bruineberg et al., 2018; Kirchhoff & Robertson, 2018; Clark, 2015). Insofar as PP is also presented as a mechanism for performing approximate Bayesian inference and has already been used to describe how we might solve a variety of "higher-level" cognitive tasks—from action-planning (Pezzulo, 2017) to counterfactual reasoning about others' mental states (Palmer et al., 2015)—so there is reason to be optimistic that it might provide an embodied framework for these "representation-hungry" forms of offline cognition that have traditionally argued to be unapproachable via sensorimotor coordination alone (Roelofs, 2018; Matthen, 2014; Clark & Toribio, 1994).

Yet, as I will argue, a recognition of the interdependence of perception and action and the rejection of a representational starting point does not an enactivist account make. The enactive approach views perception and action not only as constitutively interdependent but also as directed toward some other norm or goal. Insofar as these accounts of "sensorimotor predictive processing" do not address the question of what makes something the directed action of an agent, rather than at the mere movement of a physical object, so they constitute only a partial step toward an enactive account.

Such accounts cannot be complete unless they connect up with something like bioenactivism's grounding for the nonreconstructive normativity that we coordinate our sensorimotor engagements with respect to. Without this alternative explanation for the function of a predictive brain, Clark (2015) and others lack justification for redescribing a bare covariance between top-down and bottom-up signals in intentional terms as "prediction-errors" that a system is "trying" to minimize. Thus, PP not only stands to help scale this basic biological intentionality to "higher" forms of cognition but also stands to benefit from grounding the function of the predictive mind in the bioenactivist account of autonomy and intentionality.

So, how exactly might prediction-error minimization relate to the preservation of biological autonomy? In chapter 3, I introduce Karl Friston's theory of free energy minimization. The first component of this is active inference, a formal description of the equivalence between approximate inference and predictive control, as might both be implemented by a predictive processor. In chapter 4, I describe the second component of this, the *free energy principle* (FEP), which purports to formalize the survival of an autonomous system in terms of the kind of stability that this predictive control affords. As described in chapters 5 and 6, respectively, this formalization of inference-as-control is then supplemented with the addition of a Markov blanket, which individuates the organism from its environment, and a set of coupled stochastic differential equations, which are used to formulate the notion of a sensorimotor loop. When put together, the FEP's advocates argue that these components provide the means both to ground the intentionality of living systems *and* to scale it up to higher-level cognitive processes, via an implementation story like PP.

In viewing survival as a matter of the stability of essential variables, the FEP's account is strikingly similar to the theory of life in terms of "generalized homeostasis" offered by the cyberneticist W. R. Ashby half a century earlier. Yet, as I argue in chapter 7, biological autonomy is not reducible to homeostasis and the FEP's definition of it turns out to be trivial—as applicable to any stable mechanism as to a living organism.

Perhaps this is not the FEP's fault. Perhaps biological autonomy is not reducible because it is a vitalistic concept with no place in a good scientific account of organisms and their cognitive processes. To argue the contrary, we need to show that the FEP collapses the distinction between living organisms and machines only by neglecting essential features of the former. In chapter 8, I frame these essential features in terms of a distinction between processes and substances, to argue that the material turnover and capacity for ongoing change that are distinctive of living systems makes it impossible to capture their conditions of existence and identity in terms of the preservation of any substantial invariant features. In chapter 9, I consider a number of strategies an FEP advocate might take in order to identify some form of probabilistically describable stability amid all this turnover and change, to show that none of them succeed.

As such, I claim that the FEP is not only too general to provide an account of what distinguishes the living from the nonliving; it is also too specific in that it makes claims about necessary imperatives that simply are not necessary for organisms. Unlike inorganic structures, the ongoing existence and identity of a living system depends on neither the stability of its parts, the stability of the interactions between these, nor the stability of its overall pattern of behavior. This leaves the FEP nothing to latch on to in its attempt to define a system in terms of the preservation of a stable probabilistic model.

That the FEP fails is not enough to establish that bioenactivism has succeeded in differentiating life from non-life—in order to ascribe intentionality to the latter alone. Indeed, in chapter 10, I will argue that it has not. While the prevailing bioenactive account of autonomy in terms of process closure has advantages over the free energy framework—insofar as it provides a relational account of why some variables might need to be stabilized and why some are free to change in open-ended ways—like the FEP, it neglects the unique thermodynamic status of living things. By abstracting away from the molecular interactions of autopoiesis, in favor of closure among a network of mutually dependent and precarious processes, this “process closure” may be general enough to apply to all scales of biological organization, but it is also too general to distinguish living, intentional systems from machines.

What makes organisms special, as I describe in chapter 10, is that the apparently invariant structures that constrain and enable these precarious flows of energy (processes) are also themselves reciprocally dependent on those flows of energy in turn. It is this reciprocal dependence that Moreno and Mossio (2015) formulate in the alternative notion of constraint closure. This, I argue, succeeds in combining the advantages of both thermodynamic and relational accounts of living systems, in order to describe what it is about living systems alone that makes them genuinely autonomous agents. And it is here, rather than in the statistical constructs of the FEP, that enactivists should look to ground a teleological concept of intentionality.

1 Biodynamic Enactivism

1.1 The Enactive Approach

> The big problem I have about enactivism is figuring out what it is.
> —Ned Block (quoted in Meyer & Brancazio, 2022)

Since Varela, Thompson, and Rosch first introduced the term in *The Embodied Mind* in 1991, the "enactive approach" together with its embodied, extended, and embedded affiliates, collectively known as "4E cognitive science" has grown in popularity across philosophy, neuroscience, robotics, and AI. This swell has not amounted to a sea change, however. There may well be general agreement that we would do well to pay attention to the way cognition is sculpted by our bodies and environments, but the full-blooded enactive standpoint remains outside the mainstream.

This is perhaps unsurprising, for the phenomenological orientation of the enactive approach does not rest easily with the scientific realism that dominates both cognitive science and contemporary Anglophone philosophy. Where the scientific realist assumes investigator and investigated to be strictly independent, the phenomenologist takes self and world as inextricably entangled, views knowledge or understanding as a matter of attunement between them, and conceptualizes intentionality as the directedness of an action rather than the having of a representation. Where the scientific realist is tasked with overcoming scepticism, to explain how the deliberative operations of our internal machinery could come to mirror the independent structure of the world beyond, the problem for the phenomenologist is the question of how such a self/other distinction comes to arise at all.

As a result of this tension, much work in 4E cognition has instead adopted a scientific realist notion of embodiment in terms of a physically instantiated sensorimotor system, either seeking to extricate enactive ideas from their phenomenological frame or ignoring them entirely. Such accounts *may* follow the enactivist in rejecting a view of knowledge and cognition as directed toward the goal of accurate reconstruction. Yet, insofar as they do not supply an alternative account of the norms by which our cognitive processes are governed, they are incomplete, lacking a foundation for distinguishing a body from an object, an action from a mere movement, or an autonomous agent from a machine.

This partial uptake of a selection of enactive ideas within a scientific realist framework has led to a proliferation of distinct approaches, often with incompatible metaphysical, methodological, or epistemological commitments that either claim or have been given the "enactivist" label. Labels are a necessary evil, and the lack of a specific, widely agreed-on point of reference is often unavoidable. The same could be said of terms like "computationalist," "Christian," "fun," or "soup" (Gualeni, 2017). Still, I hope to at least be clear about what I do and do not mean by "enactivism," and "bioenactivism" specifically. I will also give some account of why I take this to pick out a coherent position and tradition within the array of ideas that are typically lumped together and presented as an alternative to "classical," "computationalist," and "cognitivist" approaches to the mind.

First, what I do and do not mean by "enactivism." One way of identifying an enactivist account is genealogical, in terms of whether it developed out of Varela et al.'s *The Embodied Mind* (1991) in which the label of "the enactive approach" was introduced. Often referred to as "autopoietic enactivism" (though as I will explain, that label is not ideal) Varela et al.'s work contains many of the ideas that I take to be central to the enactive tradition, in drawing on phenomenology to motivate the proposal of an alternative, nonrepresentationalist starting point for cognitive science. Nonetheless, others may disagree about what the "key enactivist ideas" of *The Embodied Mind* are. As such, depending on claimed connections to this particular text threatens to lump together incompatible views while excluding more closely related work on the basis of the lack of historical connections to Varela et al.'s work.

The problems with this are apparent in how the label "enactivism" is often applied quite loosely by philosophers and cognitive scientists based on one of two criteria, neither of which I take to be adequate to pick out a distinct and unified understanding of what cognition is. The first is to refer to any negative position that forbids the use of representational talk in understanding the mind (e.g., Nanay, 2014); the second, as encompassing a wide range of positive proposals, united by taking the coordination of action to play some essential, and underappreciated, role in our cognitive and perceptual lives (e.g., Gangopadhyay & Kiverstein, 2009; Ward et al., 2017).

I take identifying one's approach with the first of these to be particularly unconstructive. "Representation" means many different things to different people, and whether or not anything worthy of the name will play a role in our account of cognition is not something to be committed to at the outset but should be a downstream consequence of whatever metaphysical, conceptual, and methodological picture we develop. I take *The Embodied Mind* not as an absolute writ against any talk of "mental representations," but as presenting an argument against assuming that cognition's primary function is the veridical recapitulation of a mind-independent world, and showing how we might approach it from an alternative starting point in terms of the perceptual guidance of action.

So, the success of the enactivist program does not depend on whether it succeeds in evading representational commitments at every point of its development. Instead, the test is whether this alternative starting point does, as Varela et al. suggest, help us evade the sceptical thicket that has entangled cognitive science since its inception. Even if the enactivist is correct about this, the fact that one begins from a starting point that is not representational does not mean that one must be anti-representationalist. As Thompson (2011) remarks in relation to the emulation theory, defended in representational terms by Lucia Foglia and Rick Grush (2011): "I argue against representationalist theories that separate perception and action, instead of recognizing their constitutive interdependence, and that neglect the ways autonomous agents bring forth or enact meaning in perception and action. Since the emulation theory does not require these typical features of representationalism, my objections to representationalism need not apply to the emulation theory" (p. 19).

My intention here is not to argue that the enactive approach must *embrace* representation talk any more than it must reject it. After all, other advocates of the enactive approach advance more negative conclusions regarding the redemption of any form of representational talk (Di Paolo, 2017). I simply want to emphasize that while specific versions of the enactive approach may have found reason to reject the notion of internally realized structures in the brain that facilitate decoupled action guidance and imagination, nothing about the enactive approach in general mandates this rejection. Just because advocates of such internal structures have often chosen to call these structures "representations" or "models" does not mean that entertaining their possibility commits one to the reconstructivist, representation-first approach to the cognition that the enactive approach rejects.

While the anti-representationalist understanding of enactivism pigeonholes its development in a manner that I would reject, the second use of "enactivism" to refer to the family of views that reject the detachment of cognition and action is less pernicious. Though defining enactivism in this loose sense means grouping together a wide variety of different approaches, it does at least seem consistent with the definition of "the enactive approach" given by Varela et al. (1991) as defined by two points: "(1) perception consists in perceptually guided action, and (2) cognitive structures emerge from recurrent sensorimotor patterns that enable action to be perceptually guided" (p. 173).

This breadth can be useful, insofar as it identifies a diversity of positions that have been proposed as alternatives to what Susan Hurley (1998) called "the classical sandwich" view of cognition as a distinct, classically computationalist procedure, which occurs in between the disconnected and peripheral processes of perception and action. Still, this rejection of the "classical sandwich" is also found in the ecological psychology of J. J. Gibson (1979), the perceptual control theory of William T. Powers (1973), various "skill theories" of perception such as those of Gareth Evans (1982), Rick Grush (2007), or Susannah Schellenberg (2007), and the "interactivism" of Bickhard (2009) to name just a small scattering of examples. None of the above identifies with enactivism, and there are explicit tensions between some of these positions and the enactive view—for instance between Gibson's realism and the constructivist metaphysics of *The Embodied Mind*.

So, the view that the capacities of action, perception, and cognition are related to one another in some important way is insufficiently distinctive to identify the "enactive" view specifically. For these reasons then, I want to clearly distinguish the view that interests me from both what is sometimes called the "sensorimotor enactivism" of Kevin O'Regan and Alva Noë (2001), and from the "radical enactivism" of Dan Hutto and Erik Myin (2012). (For a nice overview of the different positions to claim the "enactive" header, see Ward et al., 2017.)

The first of these is enactivist in the second, broader sense and bears little more similarity to Varela et al.'s position than some of the "action-oriented" theories listed above. O'Regan and Noë are not so much concerned with banning representational talk but with the reconceptualization of it. As they put it, "seeing lies in the *making use* of the representation, not in the having of the representation" (2001: 1017). While sensorimotor enactivism also shares the phenomenological inspirations that animate *The Embodied Mind* and its descendants, unlike this tradition, it has tended to focus on narrower questions regarding perceptual content, rather than metaphysical and epistemological issues regarding agency, self, environment, and intentionality. That said, O'Regan and Noë's narrower proposal, which they prefer to call "sensorimotor theory," could be situated within this broader enactive project with relative ease.

The same cannot be said of Hutto and Myin's (2012, 2017) radical enactivism, which focuses more on anti-representationalism in its identification with the "enactive" label. Largely disowning the phenomenological perspective that shaped the development of the enactive approach, Hutto and Myin primarily focus on the exorcism of all talk of "intentionality" and "content" from accounts of basic cognitive processes. For Hutto and Myin, such normative attributions cannot be grounded in mere biological properties but emerge only in the level of social interactions between individuals. While I think their argument for social dependence does have plausibility when it comes to a specifically *representational* sort of normativity, the enactive approach (at least as introduced in *The Embodied Mind*) is specifically concerned with disentangling normativity from representational capacities—in order to open up the possibility of identifying a more a basic, pre-social form of normativity (Thompson, 2018). If and how enactive work can deliver this will be the topic of the next section, but for now I will merely emphasize that such a possibility is not foreclosed by argument that representation is inherently social.

1.1.1 The Enactive Approach as Naturalistic Phenomenology

So, neither self-identification nor common applications of the label "enactivist" are particularly helpful guides to picking out a coherent and distinctive approach. As such, I propose that the best way to identify enactivism is not in terms of commitment to the eradication of representational explanations in cognitive science, or as any view that analyzes cognition in terms of the coordination of perception and action, but specifically as the view that they are constitutively interdependent and intentionally directed toward a goal that is not, primarily, reconstructive. As both an inspiration for, and consequence of, this view of cognition, we find also the rejection of scientific realism and the idea that the objects of cognition and the cognizer of objects are two strictly independent realms.

Such a definition does better at distinguishing enactivism from other action-oriented approaches, such as ecological psychology and radical enactivism, while serving to identify the common thread that runs through more recent "canonically" enactivist works, such as *Mind in Life* (Thompson, 2007) and *Sensorimotor Life* (Di Paolo et al., 2017). As I will argue in the next section, it is also beneficial in allowing for the identification of shared approaches, irrespective of whether their advocates refer to particular texts or describe themselves in particular terms.

That Varela et al. (1991) took their thesis about the relation between perception and action to be more than just an empirical discovery is evident in their concern with providing not only an alternative methodology for cognitive science but also an alternative metaphysics to replace the scientific realism that was dominant at the time, and which remains so across the philosophical anglosphere today (Bourget & Chalmers, 2021). Where scientific realism takes mind and world to be strictly independent and asks how the former can become cognizant of the latter, Varela et al. explicitly situate their project within the phenomenological approach, which views mind and world, and self and environment, as inextricably entangled.

As the tradition introduced by Edmund Husserl, I understand phenomenology's central feature to be the continuation of the Kantian project to find a way between idealism and metaphysical realism. As Zahavi (2004) puts it, "Phenomenology is basically, I would insist, a transcendental philosophical endeavour, and to dismiss that part of it, is to retain something that only by equivocation can be called phenomenology" (p. 340).

Thus, while the "phenomenological method" is most commonly identified with Husserl's famous "epoché"—the practice of analyzing the objects of experience as they appear to us without distorting this through a prior commitment to their mind-independent nature—this is only the first step. The second is to attempt to identify the transcendental structures, such as perspectivity and temporality, that are preconditions for the possibility of our experiencing this world of objects at all (Zahavi, 2003; Moran, 2002).

In doing so, we can attempt to save realism about the empirical world by sacrificing a metaphysical realism that treats it as an independent "given." For the phenomenologist, this means appreciating that the world of experience is a construction that partially involves our own activity, but, crucially, that this construction is nonetheless empirically real, and not a purely subjective matter of free *individual* choice. As Varela et al. (1991) describe, "Cognition is not the representation of a pregiven world by a pregiven mind but is rather the enactment of the world and a mind on the basis of a history of the variety of actions that a being in the world performs" (p. 9).

So far, so Kantian. What differentiates phenomenology is the identification of these necessary preconditions, not with the conceptual scheme of a pure "knower" but with the "embodiment" of an agent, and consequently a developing appreciation for the contingent, dynamic, and historical nature of these transcendental structures, in contrast to the supposedly atemporal and absolute foundations of Kant's categories (Mohanty, 1978; Zahavi, 2003).

Three key points about embodiment here. First is that the term is intended to refer not only to the biological body as it is ordinarily understood but also to an extended network of cultural, linguistic, and environmental structures, insofar as these afford, solicit, and constrain possible actions. Second is that the body now becomes split accordingly into the empirical aspect given to us in experience and investigatable through scientific methods versus the transcendental lived aspect revealed via phenomenological analysis. Third, the interesting thing about linguistic, cultural, and biological structures is that they can vary and change. As such, what is revealed by phenomenological analysis as a necessary precondition for some facet of experience may, as the sociologist Alfred Schütz (1959) criticized of Husserl's early apodictic foundationalism, still be contingent on our particular situation, rather than reflective of absolute and eternal truths.

In this dual nature of the phenomenological body as Petitot et al. (1999) argue, "A transcendental analysis and a natural account are intrinsically joined." It is this recognition of the empirical aspect of transcendental structures, rather than in the discarding or downgrading of phenomenology's transcendental dimension, that Zahavi (2004) suggests opens up the possibility of a genuinely naturalistic phenomenology.

While the early Husserl may have indeed sought the "universality, necessity, apodicticity" to deliver an "absolute grounding" of human knowledge (Husserl, 1982/1913, p. 19) as Zahavi (2003) argues, he increasingly recognizes the importance of embodiment (Husserl, 2001/1920; 1997/1907), along with the revisability of phenomenological claims (Husserl, 1970/1936) and the potential for fruitful interaction between empirical and transcendental approaches (Husserl, 1999/1929). That said, it was arguably Merleau-Ponty (1964, 2012/1945) who first fully appreciated how the twofold nature of the phenomenological "body" contained the prospect for bringing together the insights of both scientific and transcendental phenomenological approaches to cognition.

As he describes it, phenomenological analysis concerns not a quest for unshakable foundations but rather "an intellectual taking over, a making explicit and clarifying of something concretely experienced" (1964, p. 68). The clear-cut distinction between such insight and an empirical fact is now blurred and recast such that, "The a priori is the fact understood, made explicit, and followed through into all the consequences of its latent logic; the a posteriori is the isolated and implicit fact" (2013/1945, p. 221). In this regard, as he argues, there is a continuity between the inductive and generalizing endeavour of the scientist and that of the phenomenologist:

> There are not two truths; there is not an inductive psychology and an intuitive philosophy. Psychological induction is never more than the methodological means of bringing to light a certain typical behaviour, and if induction includes intuition, conversely intuition does not occur in empty space. It exercises itself on the facts, on the material, on the phenomena brought to light by scientific research. There are not two kinds of knowledge, but different degrees of clarification of the same knowledge. (1964, p. 24)

The scientist and the phenomenologist share the method of attempting to extract invariant features from varying circumstances. One seeks these in the experience of bodies, languages, or societies as objects *in* our experience, the other seeks their invariant features as structures *of* experience. But neither escapes experience altogether, and neither has a direct methodological line to some pure realm of *nature* that lies outside or beyond it. Phenomenology is incompatible with scientific naturalism only insofar as the scientist forgets this and takes properties of her models to be the irrevocable truths of mind-independent reality.

As Merleau-Ponty characterizes this:

> Science manipulates things and gives up living in them. It makes its own limited models of things; operating upon these indices or variables to effect whatever transformations are permitted by their definition, it comes face to face with the

> real world only at rare intervals. Science is and always has been that admirably active, ingenious, and bold way of thinking whose fundamental bias is to treat everything as though it were an object-in-general—as though it meant nothing to us and yet was predestined for our own use. (1964, p. 290)

So, enactivism as naturalistic philosophy does not mean a natural*ized* phenomenology, wherein phenomenological descriptions of the lived body are reduced or eliminated in favor of empirical descriptions of the body as an object in our experience, but one in which phenomenological analysis and scientific method inform one another, and where neither is taken as an apodictic foundation to which the other must submit absolutely. Such a picture, as Gallagher (2018, 2017) argues, may involve not just a revision in the authority we accorded to the scientific method but also, as it does for Merleau-Ponty, a transformation in how we conceptualize its objects of investigation, toward a view of nature itself as irreducibly relational and intersubjective, constituted by the interactions between embodied agents.

It is this view of phenomenology, science, and nature, reflected in the quote from Varela et al. (1991) earlier in this section, that I take as the basis, though not the original contribution, of the enactive approach. One way of distinguishing its different strands is in terms of which dimensions of our embodiment are emphasized for investigation: whether biological self-constitution, as in Thompson (2007) or Weber and Varela (2002), sensorimotor dynamics (Di Paolo et al., 2017), or social and linguistic networks (Di Paolo et al., 2018).

While this rejection of metaphysical realism in favor of transcendental phenomenology is crucial to understanding the enactive approach, this is not to say that anything enactive must hark back to Husserl, Heidegger, Merleau-Ponty, or their direct descendants. There are other ways to the phenomenological and transcendental analysis of embodiment. This might be through an alternative post-Kantian route, for instance via Wittgenstein's similar concern with the conventional constitution of our world and the priority of our intersubjective situation and linguistic embodiment in making this possible. (For interpretations of Wittgenstein in a phenomenological light, see Overgaard, 2006; Egan et al., 2013; Gier, 1981; Zhang, 2008.)

Alternatively, one might traverse a different tradition and time period altogether: starting from the foundation of Buddhism, with Siddhārtha Gautama's distinction between ultimate and conventional truth, and following how this develops in either the Madhyamaka school, which forms the second philosophical pillar of *The Embodied Mind*, or in the Yogācāra school, where we find arguably the closest parallels with the European tradition of transcendental phenomenology (Lusthaus, 2014).

Where I take reference to the European phenomenological tradition to be particularly useful, however, is in making sense of the enactivist concern with the "intentionality" of cognition and how this sits alongside the rejection of a representationalist theory of what it is to be a cognitive system.

1.1.2 What Are Your Intentions?

The enactive approach's focus on intentional content and ideas of the body as a "vehicle of meaning" (Colombetti, 2010) has led to criticism among those who take the defining mark of enactivism to be its anti-representationalism. Hutto and Myin (2012), for instance, accuse the accounts of Varela et al. (1991), Thompson (2007), Di Paolo (2009), and Colombetti (2010) as being insufficiently radical in their continued commitment to a basic level of intentional content. This confusion about how a system can have intentional content and yet not be representational stems from the quite different ways in which the term "intentionality" is used in phenomenological and "analytic" approaches to cognition.

In both analytic philosophy of mind and classical cognitive science, the dominant notion of intentionality is as a relationship of aboutness between the content of a representational vehicle and a target object that corresponds with that content in whatever way is supposed to underpin that aboutness relationship. As such, almost all introductory textbooks emphasize at the outset that "there is no substantial philosophical link" between the philosopher's notion of intentionality and the ordinary meaning in terms of having the goal of bringing something about (Crane, 2015, p. 32). This latter state is generally explained as a subclass of the broader category of representational states, and a capacity that depends on the more basic ability to have states that are *about* things. As Crane puts it, "Intentions in the ordinary sense are intentional states, but most intentional states have little to do with intentions" (p. 32).

This view of intentionality as a relationship of aboutness toward an object, and as something more general and basic than that of an intention toward a goal, is traced to the introduction of the term into the philosophy of mind by Franz Brentano (1874). In explaining its meaning, Jacob (2019) points to the etymology of "intentionality" as deriving from the Latin "tendere" meaning to aim, strive, or tend toward.

This seems, to me, to demonstrate the very opposite of what is intended. In Jacob's example of an arrow aimed at a target, the arrow is not *about* that target. It tends toward piercing the bullseye, not becoming similar to the target or functioning as some stand-in for it. Similarly, I might aim to work harder, be more polite, get stronger, or knit faster, but in none of these cases is the aim toward some target "object" that I seek to enter a relationship of correspondence with. In ordinary English, even "object" has a second meaning tied more to purposes and goals than to aboutness, as when the detective explains that "the object of the investigation is to determine who killed Bugs Bunny," or the vice-chancellor declares that "the object of a university is to produce highly employable graduates."

As Thompson (2007) explains, where the representational theory of mind views intentionality as states having an "aboutness" relation to some mind-independent "thing," in the phenomenological tradition, intentionality is instead a property of "acts having directedness." This phenomenological intentionality still has normative content, but it is more akin to the normativity of a desire that we may succeed or fail in satisfying, rather than a depiction that may or may not be accurate.

It is precisely this notion of intentionality that is given a central role in Varela et al.'s (1991) proposal for an "enactive approach to cognitive science" (chapter 9). As they put it, "We would say that the intentionality of cognition as embodied action consists primarily in the directedness of action. Here the two-sidedness of intentionality corresponds to what the system takes its possibilities for action to be and to how the resulting situations fulfill or fail to fulfill these possibilities" (p. 206).

This directedness of an action need not, and usually does not, take the form of a deliberately formulated plan. Instead, it is something continuously manifest in our orientation to the world around us. This world, as the phenomenologist and the enactivist have it, does not first appear as a neutral array of indifferent objects about which we may later make judgements as to whether they interest us. Rather, the world, as we experience it in our unreflective engagements, appears a landscape of possibilities for action that may solicit or repel us. This is nicely described in a classic example from Merleau-Ponty's *The Structure of Behavior* (1963):

> For the player in action the football field is not an “object,” that is, the ideal term which can give rise to an indefinite multiplicity of perspectival views and remain equivalent under its apparent transformations. It is pervaded with lines of force (the “yard lines”; those which demarcate the “penalty area”) and articulated in sectors (for example, the “openings” between the adversaries) which call for a certain mode of action and which initiate and guide the action as if the player were unaware of it. The field itself is not given to him, but present as the immanent term of his practical intentions; the player becomes one with it and feels the direction of the “goal,” for example, just as immediately as the vertical and the horizontal planes of his own body. (p. 168)

From phenomenology, we thus gain an alternative account of cognition, knowledge, or understanding. One that does not take these, primarily, in terms of having some internal representation with depictive content that is evaluated in terms of whether it accurately corresponds with an independent state of affairs. Instead, cognition is to be explained in terms of our skill of appropriately responding to the solicitations and affordances of our surroundings, as illustrated in the practical knowledge of a typist, musician, or sportsperson, an idea picked up by Dreyfus (2002) as “skilled coping,” or, in sensorimotor theory, the notion of “sensorimotor mastery” (O’Regan & Noë, 2001).

While this phrasing is all very nice in moving toward a nonrepresentational framing of intentionality, the notions of skill or mastery imply not only appreciation for what is possible but discernment as to what is preferable. The perceptual world for Merleau-Ponty is not merely a neutral matrix of “I cans.” It is, as described above, an affective milieu, a field of salience and significance with lines of force that draw and repel us. If cognition consists in our skill at responding to these forces to attune with the world, or to increase our grip on it, then what is the standard by which this attunement is judged?

In Merleau-Ponty, as in Husserl, this normativity is often characterized in terms of epistemic exploration. Husserl (2001/1920) rather sensuously describes the unseen parts of an object as something that “calls out to us, as it were, in these referential implications: ‘There is still more to see here, turn me so you can see all my sides, let your gaze run through me, draw closer to me, open me up, divide me up; keep on looking me over again and again, turning me to see all sides’” (p. 41).

Merleau-Ponty (2012/1945) similarly speaks of being drawn toward the optimal viewpoint for a painting in an art gallery, or the understanding of how to shift an object in relation to background lighting in order to best discern its color. Yet, Merleau-Ponty, if not Husserl, recognizes that there is more to the norms of attunement than improving one's epistemic standing. Football is not an epistemic activity. The player on the field is not drawn to intercept just to learn what it feels like, but because the goal of the game is to score more goals and their role as a defender is to prevent the other side from scoring. Likewise, the classic examples of typing, organ-playing, dancing, and climbing are not purely activities of exploration and discovery; they are governed by other norms, of linguistic coherence, elegance, or ascension. As such, in his analysis of the forms of our experience in *The Structure of Behaviour*, Merleau-Ponty is concerned also with the structures of what he terms the "vital order" and the "human order."

As he describes with respect to the former:

> Thus each organism, in the presence of a given milieu, has its optimal conditions of activity and its proper manner of realizing equilibrium; and the internal determinants of this equilibrium are not given by a plurality of vectors, but by a general attitude toward the world. This is the reason why inorganic structures can be expressed by a law while organic structures are understood only by a norm, by a certain type of transitive action which characterizes the individual. (1963/1942, p. 148)

While these allow us to draw on an array of recognized, non-epistemic norms—from standing the appropriate distance from other people to running away from danger—merely pointing to these norms does not explain their origin and force. This is a problem still faced by O'Regan and Noë's (2001) sensorimotor theory, which, in lacking a theory of selfhood and autonomous agency, does not have the tools to move beyond the mere awareness of neutral sensorimotor possibilities to explain the "affective allure" in how particular affordances "grab" or solicit us.

For the enactive approach then, the project of naturalizing intentional content and that of grounding teleology, or normativity, are intertwined. In all cases, what we need is an account of what it means for some activity to be directed toward an end, in a manner such that it can be described as genuinely "trying" to achieve that end, with the consequent possibility of failing. Moreover, a satisfactory account must neither default to an unsupported standard of accurate representation nor enter into a regress by calling on the further intentions and projections of some external designer. Just as different strands of the enactive approach may focus on different aspects of embodiment, so they may look to correspondingly different places to ground this teleology.

Situating the enactive approach as a continuation of naturalistic phenomenology—that is to say, as a foundational enquiry into the nature of knowledge and reality, rather than as a local theory about the objects and methodology of cognitive science specifically—is a better way to identify it as a coherent program. This not only clearly distinguishes the enactive approach from other more local accounts that do not share these aims, such as radical enactivism or sensorimotor theory, but, as I will argue, allows us to identify other instances of the same approach, irrespective of either terminological choices or direct historical links. As with transcendental phenomenology in general, there are other routes to the same ideas. After all, if this is indeed a promising approach, then it would be strange if no one else had hit upon it.

1.1.3 Hurley's Enactive Approach

So, the phenomenological inspiration behind the development of the enactive approach is essential to understanding the particular way in which Varela, Thompson, and Rosch describe their project, and how it differs from other views described as "enactive." As I mentioned, however, there are alternative routes to the same view of perception and action as constitutively interdependent and directed toward a goal that is not, primarily, reconstructive. Of particular note is the view of Susan Hurley, who, while beginning with Kant, takes a different route via Wittgenstein, rather than Husserl, Heidegger, or Merleau-Ponty, to develop a view that I take to be far closer to that of Varela et al. (1991) than more commonly cited examples of the enactive tradition.

Hurley's does not claim the appellation enactivist, nor does she draw significantly on Varela et al.'s (1991) work—while her book, *Consciousness in Action,* was not published until 1998, the acknowledgements note that it was written just a year prior to the publication of *The Embodied Mind*. Nonetheless, she is occasionally classed as a sensorimotor enactivist—a classification that, as Ward (2016) demonstrates, does a disservice to the sophistication of her account. Rather than just taking the contents of perception to depend on our knowledge of how our movement will change our sensory input, as O'Regan and Noë (2001) do, Hurley proposes a "two-level *interdependence* view, whereby at both the subpersonal level of sensory inputs and motor outputs, and at the personal level of perception and agency, the capacity to act and the capacity to sense are necessary preconditions for the possibility of each other" (Hurley, 1998).

The emphasis on interdependence distinguishes Hurley's enactivism from the various skill theories of perception or the control theories of action. For Hurley, it is essential too that this interdependence is not merely instrumental, as in Gibson's emphasis on their utility with respect to each other's independent functioning, but a constitutive matter of what perception and action *are*. Action just is the control of perception, and perception just is the presentation of possibilities for action. As such, they are necessarily inseparable. In this sense, Ward (2016) argues, Hurley is best described as a "transcendental enactivist."

In her shared circuits model, Hurley (2008) proposes how this might be scaled up through a hierarchical structure of control systems, a striking precursor to the predictive processing accounts I will discuss in the next chapter. As has been suggested as a consequence of predictive processing (Kiefer & Hohwy, 2018), so Hurley also proposes that cognitive contents must thus be attributed holistically in terms of the potentially flexible relations between input and output and the perceptions and intentions that make up this overall control system.

This leaves us with the question, "Control in the service of what?" Hurley is particularly sensitive to the threat of what she terms "the myth of the giving," whereby one proposes to explain the content of a perception via the content of an intention, which is taken to somehow be primitive and in need of no further explanation. In itself, such a strategy is no better than treating the objects presented in perception as the immediate "givens" of a mind-independent world. Attempting to avoid the subjective regress of this type of "just more content" strategy, Hurley argues that we need a replacement for the role played by the representation of an external world in delivering a nonsubjective grounding for content determination. This, she proposes, is what an objective account of normativity could deliver.

The problem, as Hurley (2003) describes, is in distinguishing a system that is genuinely following a norm at which it might fail from the operation of a simple feedback control system, like a programmable thermostat, that also adjusts internal connections between input and output. The thermostat may fail according to *our* goals, but there is nothing to prevent describing its behavior as successfully following some alternative rule. True normativity, she suggests, depends firstly on something like increased context-sensitivity and flexibility in how a goal is pursued, of the sort afforded by the higher levels of her shared control circuits model, and secondly on some external teleological constraints such as social context and evolutionary pressures.

As Hurley (1998, 2003) notes, however, she lacks an account of exactly what grounds this supposed "teleological context" and distinguishes it from the basic physical laws that apply to agents and nonintentional systems alike. Mere complexity of behavior does not appear adequate to the task, for no amount of complexification seems sufficient to entirely dispel the concern that a robot capable of following (or at least appearing to follow) some sophisticated array of norms might nevertheless be as mindless a zombie as any other machine. As she puts it:

> An agent with conceptual abilities has a more richly structured set of behaviors, and perhaps those behaviors must have causes with a certain related structure. But even granting all this, it is not clear why machine or zombie worries should be disarmed by conceptual abilities. If these worries are valid in the first place, why couldn't a machine or zombie have a conceptually structured set of behaviors and reasoning abilities with correspondingly structured causes, yet not be in conscious states? If the worries get a grip to begin with, their grip is not loosened by the enrichment of structure and of norms of rational behavior that goes with conceptual abilities. (1998, p. 162)

Perhaps, she goes on to suggest, "the extra ingredient needed in a set of sufficient conditions is not the richer normativity of conceptual abilities, but simply life" (p. 162). Yet, she does not pursue this possibility. As indicated, her concern is with the gap between a norm-following intentional agent and a conscious one, a concern in which she simply presumes a double disassociability between intentional action and life.

On the bioenactive view I will develop in this book, life and intentionality are not so easily separated. While it might be possible to create life without agency, once we see what life has to contribute to the naturalization of the normative domain, we will see why there can be no intentional agency without it. Perhaps, though it is not a topic pursued here, once we have such a rich account of intentional agency, any further puzzle about whether additional ingredients are needed for consciousness will dissolve.

This problem of how normativity and intentionality emerge is not resolved in *The Embodied Mind* either. Despite being explicitly concerned with resisting the elimination of subjectivity, intentionality, and agency in favor of the mere objects of the physical sciences, Varela et al. do not go beyond the explicitly mechanistic notion of autonomy as operational closure in the attempt to characterize these. Crucially, they suggest that the principal difference between their simplified illustration of this in the cellular automata, Bittorio, and a living organism is simply in their respective degrees of complexity. I disagree. As I will argue, this operational closure alone fails to capture the difference in kind between the intentional and teleologically oriented agent that the enactivist needs versus a mere physical mechanism.

The development of the bioenactive approach thus begins not with *The Embodied Mind* but with the notion that we can ground these teleological and intentional properties in the self-production of an organism, specifically its intrinsic dependence on an internal metabolic network. While the idea of metabolism as teleological originates with Hans Jonas (1953, 2001/1966), the uptake of this by enactivism can be traced to Weber and Varela (2002) and is continued through Thompson's (2007) *Mind in Life*. So, while enactivism is distinguished by a view of cognition as the coordination of perception and action, where these are constitutively interdependent directed toward a goal that is not, primarily, reconstructive, *bioenactivism*, as I will now describe, is the attempt to ground this intentional directedness toward a goal in biological terms.

1.2 Bioenactivism

The bioenactivist inherits a commitment to both naturalism and realism about teleology and intentionality from enactivism more generally and adds to these two further commitments: mind-life continuity and the view of life as self-production. These are supposed to describe how our biological embodiment accounts for the normative and teleological dimensions that the enactive approach posits but does not explain.

Mind-life continuity is the simpler of these to explain and is expressed nicely by Hans Jonas (2001/1966) as the view that "The organic even in its lowest forms prefigures mind, and that mind even on its highest reaches remains part of the organic" (p. 1). How such a claim is interpreted will naturally depend on what one takes as the relevant features that both share. Life-mind continuity is thus a relatively minimal commitment shared by a wide diversity of theorists whom Lyon (2006) surveys as the various "biogenic approaches to cognition" (p. 11).

If we, as enactivists, take the "mark of the mental" to be its intentional directedness (non-representationally understood) then mind-life continuity means locating this intentional directedness in some property of living systems. As Thompson (2007) notes, it is this focus on the existential and phenomenological dimensions shared by the biological and the mental that differentiates the Jonasian and enactive approach from other life-mind continuity approaches, which focus on shared organizational or functional aspects.

When it comes to accounting for the appearance of teleology and intentionality in particular, the standard recourse since Darwin has been to natural selection, to explain the development of increasingly complex forms of organization in terms of heritable variation and differential reproductive fitness. Yet, Darwin's achievement is often viewed not as making room for intentionality in our naturalistic worldview, but rather as eliminating it (Stenmark, 2001). As Dawkins (1986) puts it, "Natural selection, the blind, unconscious, automatic process which Darwin discovered, and which we now know is the explanation for the existence and apparently purposeful form of all life, has no purpose in mind" (p. 5). And, as he writes elsewhere, "The universe we observe has precisely the properties we should expect if there is, at bottom, no design, no purpose, no evil and no good, nothing but blind, pitiless indifference" (1995, p. 133).

We might attempt to retain a teleological dimension in the tendency toward increased reproductive fitness described by natural selection—for instance by pointing to how, unlike the exceptionless laws of physics, this can "fail" to be realized in particular individual cases. The same is true of the second law of thermodynamics (Wicken, 1981). In both cases, however, this might be better accounted for by viewing these as statistical generalizations from underlying causal processes, rather than taking them as laws themselves—teleological or otherwise (Matthen & Ariew, 2002). Moreover, even if increased entropy or reproductive fitness were indeed a purpose toward which the universe is being driven, in either case, this could not be credited to the work of individual agents, trying or failing to follow that norm.

Accordingly, bioenactivism has taken a different route, looking instead to the level of the individual organism and the basic unit of organic life in the single cell. Thus, while its philosophical roots may be in phenomenology, its biological ones are in autopoiesis theory, developed by Humberto Maturana and Francisco Varela, which proposes to identify the basic logic of this self-production in the single cell, in order to then formulate its essential features in the more general notion of autonomy.

1.2.1 Autopoesis and Autonomy

For Maturana and Varela (1973/1980) an autopoietic system is defined as such:

> An autopoietic machine is a machine organized (defined as a unity) as a network of processes of production (transformation and destruction) of components that produces components which:
>
> (i) through their interactions and transformations continuously regenerate and realize the network of processes (relations) that produce them; and
>
> (ii) constitute it (the machine) as a concrete unity in the space in which they (the components) exist by specifying the topological domain of its realization as a network. (Maturana & Varela 1980, 78–79)

As this is realized in a cell, the relevant aspects are a membrane or boundary separating the interior from the external environment and, inside this boundary, a network of enzymes and reactants fed by those molecules that are allowed in through the membrane, which either generate further reactants and enzymes, or the components making up the cell's boundary. In this respect, as Boden (2000) describes, autopoiesis can be viewed as an attempt to define the metabolism of a cell in organizational terms, thereby allowing us to abstract away from particular chemical components.

Because of the use of cellular autopoiesis as an illustrative example, what I call "bioenactivism" here is often referred to as "autopoietic enactivism." While this is perhaps a better label insofar as it flags which particular sorts of biological properties the enactivist is concerned with, this name is, as previously mentioned, misleading for at least two reasons.

The first reason is that, as Thompson and Di Paolo (2014) point out, the crucial concept in *The Embodied Mind* was not autopoiesis specifically, but rather the more general principle of *autonomy*—of which cellular autopoiesis is related to as an instantiation at the molecular level. Autonomous systems are defined as networks of processes that exhibit *operational closure:* a recurrent organization such that each process in the network both enables another, and depends on another in turn, and *precariousness*: such that were any of these processes to break down, the network as a whole would cease to exist (Thompson & Di Paolo, 2014). It is by means of these properties of recurrent organization and mutual dependence—rather than in molecular reactions and membranes—that a living system distinguishes itself from its environment, as shown in figure 1.1.

This is then supposed to give us normativity, insofar as the operationally closed organization of a system defines a domain of interactions compatible with the maintenance of this precarious system. Thus, as Thompson (2007) puts it, "Cognition is behavior or conduct in relation to meaning and norms that the system itself enacts or brings forth on the basis of its autonomy" (p. 158).

Cellular autopoiesis is thus to autonomy rather like what Turing's (1948) description of his tape-based machine is to computation, and referring to the enactive approach as "autopoietic enactivism" is a bit like referring to computational accounts of the mind as "paper-and-pen cognitivism." It

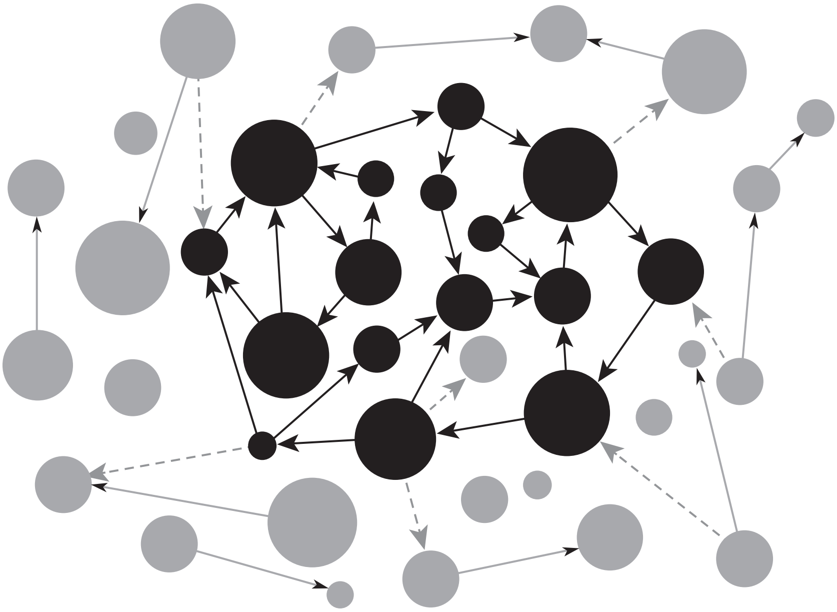


Figure 1.1

An illustration of how an operationally closed network of processes (highlighted in black) is distinguished from its surroundings. One-way dependencies of either enablement or dependence alone do not admit membership of said network and are shown with dashed lines (Di Paolo & Thompson, 2014).

is the notion of autonomy that is supposed to provide a characterization of the logic of living systems that can be generalized across varying levels of organization from the single cell up to the recurrent dynamics of the nervous system.

A second reason the emphasis on autopoiesis can be unhelpful is because, as Villalobos (2013) and Villalobos and Ward (2015) describe, it encourages the muddling together of autopoiesis theory and bioenactivism. While the concepts of autopoiesis and autonomy were both developed in Varela's work with Maturana, the latter is adamantly not an enactivist—nor would Varela have been understood as such throughout much of their collaboration. Maturana's goal is not to differentiate living systems from machines but to explain how a machine could be alive. Thus, as he states in his introduction to their coauthored *Autopoiesis and Cognition: The Realization of the Living*, in characterizing living systems, "notions of purpose, goal, use or function, had to be rejected" (1972/1980, pxiii).

In this respect, and in the aim to reduce intentional talk to the operations of feedback control systems, autopoiesis theory is cybernetic, not bioenactivist. In the spirit of the British cybernetician W. R. Ashby, it endorses a form of life-mind continuity to view the operations of our mind as an elaboration of the homeostatic regulation of simpler organisms. Unlike bioenactivism, however, it follows this continuity all the way into nonliving matter, taking there to be no difference in kind between the existential status of an organism and any other physical system, and thus no basis for attributing intentionality or intrinsic purposiveness as a real property that is uniquely possessed by the latter.

Enactivism itself is not introduced until *The Embodied Mind*, which, as noted, does not yet tackle the problem of naturalizing teleology head-on. It is only by the late 1990s, where Varela has been influenced by Kant and Jonas's work on the idea of organisms as "natural purposes" that he begins exploring the concepts of "original intentionality" and "sense-making" as unique to life, coming around to the position that these do lead to the reintroduction of a kind of teleology that is "intrinsic to life in action" (quoted from an email exchange in Thompson, 2007, p. 454). This culminates in a 2002 article with Andreas Weber that draws on Jonas's (2001/1966) attempt to naturalize teleology in the 'needful freedom' of a metabolic system.

This "needful freedom" is intended to emphasize that a metabolic system is not just a machine that can freely persist through a turnover of material components—as implied by Maturana and Varela's earlier claim that "autopoietic systems are homeostatic systems which have their own organization as the variable that they maintain constant" (1972/1980, p. 80). What makes a metabolic system "needful," rather than merely "free," is that it is *dependent* on this material turnover and its own synthesizing activity for its continued existence. A chair need do nothing at all in order to carry on being a chair, but it is not merely a human projection to say that if a cell's metabolic activity breaks down, then the cell breaks down along with it. Part of what it is to be a cell, part of what it is to be a living thing, is to be something that works toward its own ongoing production through the continual turnover of molecular material. This, Jonas suggests, gives a purposiveness dimension to the cell's activity, such that we can describe its breakdown as "failure" even where this is a deterministic outcome inevitably entailed by some prior event.

It is in this definitive break with Maturana's insistence on treating both living and nonliving systems as purposeless mechanisms alike that I would locate the origin of bioenactivism. Two vital questions for its development are: firstly, whether the prevailing definition of autonomy as operational closure among precarious processes captures the needful freedom found in the metabolic cell, and secondly, whether this definition of autonomy is adequate to ground the kind of purposive and intentional attributes needed for an enactive account of cognition. As I will argue, the answer to the first question is negative, and, as a result, bioenactivism has thus failed to provide a positive solution to the second question. This does not mean there is no such solution, however, but only that bioenactivists need a better formulation of autonomy. Just such an account, as I will argue in chapter 11, is provided by Montèvil and Mossio (2015) and Mossio and Moreno's (2015) account in terms of "constraint closure."

Before we get to these issues, however, I want to look at a second development in the bioenactive literature, stemming from a different sort of dissatisfaction with how prior formulations of autonomy and autopoiesis are supposed to connect up to enactive norms. This is the fact that, as Di Paolo (2005) notes, the conservation of autopoietic or autonomous network gives us only the all-or-nothing norm of "don't die" and "the rather useless *a posteriori* realization by the external observer that the organism should have avoided that very last encounter that killed it" (p. 436). The bioenactivist needs more: namely, an account of graded norms that an agent can work toward and that can guide its sensorimotor interactions toward increasing attunement.

1.2.2 Sense-Making and Adaptivity

In addition to realizing its own ongoing existence, an organism's autonomous organization also implies a window of viability—a specific range of environmental conditions outside of which the processes making up the autonomous system will break down. To take cellular autopoiesis: this requires both particular states of affairs (e.g., temperature and pressure) and the ongoing supply of the necessary components to fuel the cell's metabolic processes. As our planet is not a homogenous sphere of lukewarm nutrient soup, so even the simplest of living systems must also adjust to, and interact with, its environment in order to maintain itself within this window of viability.

While the importance of world-engaging sensorimotor patterns in constituting a subjective perspective was central in *The Embodied Mind* (as reflected in the quote cited in section 1.1) the failure to adequately integrate this with the autopoietic-autonomous constitution of the bounded individual reflects a conceptual tension that Barandiaran (2017) argues has troubled the enactive approach from the beginning—a tension between the organism as separated from the environment, and yet also defined by and dependent on its interactions with it (Bitbol & Luisi, 2004; Bourgine & Stewart, 2004).

This tension is explicit in the separate development of sensorimotor enactivism, as a description of the content and structure of perceptual experience in terms of the dynamic relationships between sensory input and motor output. While such an account shares in the broader enactivist rejection of reconstructivism, it lacks foundation in an account of the norms by which these sensorimotor interactions are coordinated. On the other hand, the concepts of autopoiesis and autonomy in isolation are insufficient to capture the logic of cognition. While they provide us with a naturalistically founded "basic" normativity, it is only the all-or-nothing imperative of ongoing self-production. We have our success criteria—the preservation of the autonomous network that constitutes the organism—but we need also the criteria that a living system directs its environmental interactions toward that end.

Let's take a simple example of apparently norm-governed behavior: bacterial chemotaxis. A favored case study of Varela's (1991) and now a mainstay of biogenic approaches to cognition (Lyon, 2006), this describes how a bacterium controls the motion of its flagella in order to move toward higher concentrations of glucose. At a minimum, doing so involves sensing of current glucose concentration, the memory of previous concentration levels, the comparison of the two and the activation of an appropriate motoric response—initiating flagella-rotation to switch from directionless flailing to a directed run when concentration increases. This is a minimal example of "intentional" behavior without a reconstructive model. There is no internal "utility heatmap" inside the bacterium that represents the distribution of nutrients throughout the current solution, by means of which it plans its journey. There are simply a series of sensorimotor connections of the form—"If an increase in sugar concentration is detected, then engage flagella rotation."

So, bacteria swim toward sugar. But, before we get carried away and excitedly attribute intentionality, purposiveness, cognition, and subjectivity to this process, it should also be pointed out that rocks fall toward the ground—but no one is inclined to claim that they intend to do so. It is not simply the fact the bacteria reliably swim toward sugar that legitimates the enactivist attribution of intentionality to this behavior, but the metabolically underwritten fact that if a bacterium does *not* swim toward sugar, it is unlikely to remain a bacterium much longer. The glucose has "significance" for the bacterium as a nutrient, but this significance is not reducible to the physical properties of the glucose alone. This significance can only be understood in terms of the relationship between the bacterium-self and the sugar-world. It is this organism-environment relationship that brings forth, enacts, or constitutes a phenomenal world of significance valence—a world that, as Merleau-Ponty (1963) argued, is irreducible to either the flagella-rotating action of our independent subject or the chemical properties of the metaphysically independent object.

For the sense-making, or enaction, that constitutes this phenomenological, relational world then, we need both autonomy and the capacity to interact in service of its preservation.[1] Di Paolo (2005) addresses this latter requirement under the heading of "adaptivity" and defines it as such:

> A system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,
>
> 1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,

> 2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second. (Di Paolo, 2005, p. 438)

Organisms not only engage in autonomous self-production but also monitor and regulate their internal states and environmental interactions in order to remain in conditions where this self-production is possible. The similarity between adaptivity and the notion of homeostasis is immediately apparent. Yet, just as autopoiesis implies more than metabolism, so too does adaptivity involve more than mere homeostasis. For while homeostasis refers to the preservation of "key variables," such as body temperature, within particular bounds, the conservation of adaptivity is about the conservation of overall organization. Homeostasis is a derivative of this more fundamental requirement. Secondly, while examples of homeostasis typically focus on internal regulatory processes that act directly on these variables, such as osmoregulation, Di Paolo's definition of adaptivity explicitly emphasizes regulation of the organism's relation to the environment also—that is, regulation via the kind of extended agent-environment loops found in bacterial chemotaxis.

Moreover, as Di Paolo develops his account, adaptivity does not merely involve the activation of particular physiological or sensorimotor processes, to regulate internal states, but also the regulation of these processes themselves in response to environmental change. This goes beyond consistent chemotaxis to one particular nutrient source, a distinction nicely illustrated by the example of the Lac-Operon mechanism in *E. coli* (Jacob & Monod, 1961). As Di Paolo et al. (2017) describe:

> Under normal conditions *E. coli* metabolizes glucose. But when availability of this sugar is low while another (lactose) is abundant, certain normally inactive genes will be expressed that enable a new metabolic path way allowing for the processing of the new sugar. In effect, the bacterium detects a change in environmental conditions that jeopardizes its self-maintenance and reacts by modifying the internal processes underlying its self-construction. It is easier here to distinguish the regulation of behaviour from its normal execution, as the normally dormant genes that are activated contingent on specific environmental conditions do not take part in the ongoing self-sustaining processes of the organism. (p. 130)

So, without the graded normativity of adaptivity, there is no sense-making, no phenomenal world of affective forces to draw and repel us between better or worse modes of interaction with the world. As such, in their book *Sensorimotor Life*, Di Paolo et al. (2017) take adaptivity, alongside the self-individuation of autopoiesis, as a necessary requirement for the attribution of agency and subjectivity to living systems.

1.3 From Bioenactivism to Predictive Processing

There is still, however, something missing in Di Paolo's (2005) account of adaptivity. Or at least something glossed over in the move from the regulation of adaptive processes to talk of the "experience-dependent" discovery of new ones during individual development.

As in the *E. coli*, natural selection over multiple generations can lead to the development and subsequent inheritance of various adaptive responses—not only to directly threatening changes to homeostatic variables themselves but also anticipatory adjustments in response to signs, such as predator footprints, that act as a proxy for an imminent threat to those variables. Such mechanisms, if not always successful, remain relevant across the variety of environments the individuals of a particular species may have to contend with. But this lesson had to be learned the hard way, at the level of the species, by means of the breakdown in autonomy for any individual that failed to successfully inherit and activate them.

Neither the ability to switch between different genetically encoded, responses nor the ability to activate these in anticipation of a non-immanent threat, amounts to the experience-dependent learning of new ones. The question is how the individual organism can learn that a novel signal, say the rising smell of sulfur, threatens its continued existence, without this coming, as it were, one lesson too late? How does an individual learn new adaptive responses to such novel threats without actually experiencing its own breakdown?

In *Sensorimotor Life*, Di Paolo et al. (2017) draw on Piaget's detailed account of sensorimotor equilibration to describe how a system assimilates new environmental possibilities and accommodates these via alterations in its sensorimotor organization. Yet, as they state, their presentation is advanced "at the meso level between neurodynamics and personal experience" and as such, does not say "much about how sensorimotor scheme equilibration and selection is supported in the brain and other body structures" (p. 105) As they claim,

> The dynamical systems approach to sensorimotor equilibration is not a fully developed theory. It outlines the essential elements that such a theory will eventually

> have to contain, but several details, for example, regarding its possible implementations, have yet to be filled in. Progress in this area will need to involve further work on the nature of open-ended learning: for instance, additional examination of the processes assumed to be open-ended in nature (such as biological evolution and the dynamics of immune networks) and their relation to processes that could be operating in the brain (see, e.g., Fernando et al. 2012; Watson and Szathmáry 2016) and in the non-neural body. (p. 105–106)

Di Paolo et al. (2017) are skeptical about the utility of predictive processing accounts for playing this role, seeing them as unduly internalistic, representational, and committed to a priority of perception over action. While predictive processing has often been described in such terms a number of authors have argued that these are not essential commitments of the framework (Vázquez, 2020; Nave et al., 2020; Harvey, 2018; Bruineberg et al., 2018; Kirchhoff & Robertson, 2018; Clark, 2015). There is no in-principle reason why enactive goals could not sometimes be served by predictive means. As such, rather than rejecting predictivist accounts of "higher" cognitive activities, such as dreaming (Windt, 2018), action-planning (Pezzulo, 2017), memory (Henson & Gagnepain, 2010), and counterfactual reasoning about others' mental states (Palmer et al., 2015), I believe we should look to see whether the mechanisms described might be recast in terms of sensorimotor scheme equilibration and selection.

The possibility that predictive mechanisms may equally be redescribed as in the business of sensorimotor coordination, the business of "delivering a grip on the *patterns that matter* for the *interactions that matter*" (Clark, 2015, p. 19) not only opens up the possibility for PP accounts to be appropriated by enactive cognitive science but highlights the extent to which PP is itself in need of a bioenactive account of intentionality. The interpretation of some operation (neural or otherwise) in terms of anticipatory content cannot just be read off of its structure and dynamics. Like any attribution of content, representational or not, this is dependent on norms of functionality, such that we can talk of these "anticipations" as succeeding or failing (Millikan, 1984; Hurley, 1998). The redescription of PP in enactive terms merely rejects the proposal of a *reconstructive* function but does not provide the needed alternative.

Thus, to my mind, the most important objection that Di Paolo et al. (2017) raise to predictivist frameworks does not rest on whether they are described in reconstructivist or sensorimotor terms. Rather, the central issue is that the appropriateness of normative or agential terms cannot merely be read of the dynamics of the system alone. The predictivist framework in itself has no account of what distinguishes the sensorimotor dynamics of an agent from those of a mere control system, such as a thermostat or guided missile. This echoes the problem with Noë and O'Regan's (2001) sensorimotor theory, discussed in section 1.1.2—namely, that neither account has a basis for attributing normative dimensions like "grip," "attunement," or "sensorimotor mastery" to our interactions with the world.

It is with a view to describing the constitution of an agent, and to addressing these normative dimensions, that predictive processing is often motivated via Karl Friston's free energy principle. In contrast to the above claim that mere prediction-error-minimizing dynamics cannot supply an account of agency, Friston and colleagues explicitly assert that an imperative toward the minimization free energy (roughly, a generalized notion of prediction-error) captures, as they put it, "the 'intentionality' or 'aboutness' of living systems—that is, the directedness of the organism towards a meaningful world of significance and valence" (Ramstead et al., 2018, p. 33). Moreover, this principle, they propose, can "supersede or absorb classical (i.e., autopoietic) formulations of enactivism" (Ramstead et al. 2021, p. 59).

The core argument of this book is to show how such a claim fundamentally misconstrues autopoiesis, autonomy, enactivism, and indeed life in general. Still, the fact that free energy minimization fails as a formulation of autonomy does *not*, however, mean that free energy or prediction-error minimization accounts have nothing to offer the bioenactivist. Once we have both the details of predictive processing and the free energy principle on the table (chapters 2–7), and a better understanding of how autonomy is exhibited in biological systems (chapters 10–11), then we will be able to see both the extent to which the coordination of adaptive actions for the preservation of autonomy may sometimes be served by prediction-error minimization and the degree to which the relation between what is adaptable and what is predictable will only ever be contingent and approximate.

But before we get to the disputed terrain of intentionality, autonomy, and bioenactivism, let's set some ground rules. Where did the predictive processing theory of cognition come from, and what are the structural constraints that both the reconstructivist and the enactivist can agree on as entailed by (though not sufficient for) the claim that a system is a predictive processor?

2 Predictive Processing

2.1 Minimal Predictive Processing

The introduction of predictive processing (PP) in the philosophy of mind and cognitive science can be traced to the work of Jakob Hohwy (2013) and Andy Clark (2013, 2016), both of whom pick up Karl Friston's (2003, 2005, 2010) proposal for how predictive coding might be used to provide a generalized theory of brain functioning (2005), said proposal, in turn, having been influenced by Rao and Ballard's (1999) account of hierarchical predictive coding in the visual cortex.

While predictive coding is the most basic component of PP, it is not distinct to PP—indeed its origins are not in neuroscience but data compression, as a strategy developed as a means for the storage and transmission of image and video files during the 1950s (Clark, 2016; for an overview, see Shi & Sun, 1999; Musmann, 1979). The basic idea is that there are typically regular patterns in the data that we wish to store or send, and, thus, rather than encoding the value of each pixel individually, we can encode an image more efficiently by only encoding this pattern and its occasional violations. In a video, for instance, large areas of background often remain unchanged over some duration, so, rather than retransmitting the entire scene anew for each frame, we can simply transmit the pattern once, then only encode the subsequent "errors" induced by local movements of objects and agents in front of this background.

The proposal that the brain uses such a coding strategy, as Sprevak (2021) describes, dates back at least to Attneave (1954) and Barlow (1961), who argued that bottlenecks in the early visual system—for instance, the number of neurons, their dynamic range, limitations on firing rate, and the metabolic costs of firing—require the brain to use such "redundancy-reducing" code for the transmission of sensory data (see also Zhaoping [2006] for a review of some relevant constraints).

This already introduces some minimal requirements on the structure of the brain: namely, a distinction between prediction neurons (sometimes given the theoretically overloaded name of "representation neurons") and a comparator or prediction-error neuron, with signals flowing both "downward" and "upward" between these. A second source of input to the comparator neurons is from incoming sensory signals, which are compared to the downward prediction signal. The signals from the prediction neurons continue to change until they match the sensory input, signaling that it has been effectively "predicted" (Keller & Mrsic-Flogel, 2018).

Hierarchical predictive processing adds to this a couple of specific proposals about how predictive coding is implemented in the brain, with consequent requirements on the architecture of a system that could qualify as a predictive processor.

These are:

1. **Hierarchy:** This process is repeated at various levels, where the input to one level is the state of the level directly below, bottoming out in the sensory periphery.
2. **Precision-weighting:** Predictions and prediction errors are assigned a relative weight, corresponding to the inverse variance of the signal, that determines the influence a prediction error has in changing a prediction.

The hierarchical aspect means that only the bottom level is concerned with matching the sensory signal directly, with each ascending level being driven by the prediction of regularities over increasing spatiotemporal scales. In such a hierarchy, there are many degrees of freedom as to which prediction neurons should adjust in order to match incoming signals; thus, the role of precision weighting is to determine where this adjustment happens—namely, in those neurons where the prediction error signal has a high precision-weighting relative to that of the prediction neuron.

The origins of this account are typically traced to Rao and Ballard (1999), who showed that taking this as a model of the visual cortex predicted a variety of known neural responses, such as end-stopping, not attributable to classic receptive field effects alone. Friston (2005) then extended this to the whole cortex to show how it accounts for a variety of further empirical predictions concerning anatomy and synaptic plasticity; electrophysiological effects, such as mismatch negativity; and psychophysical ones, such as global precedence and priming.

This is not yet predictive processing, which you may be familiar with as "the emerging unifying vision of the brain as an organ of prediction using a hierarchical generative model." (Clark, 2013, p. 5). Having just read a section on enactive approaches, the missing piece should be obvious—the brain is not just a perceiver but an actor. It is the addition of action into the predictive processing story that is arguably the main feature that has come to distinguish Friston and colleagues' work on general predictive accounts of the brain, under the heading of "active inference" (Friston, 2003, 2010; Brown et al., 2011), and it is this that is crucial to Hohwy, Clark, and subsequent philosophers' discussions of predictive processing. On these accounts, prediction error relative to expected sensory input may not only drive internal changes to prediction neurons but can, alternatively, drive actions that reduce this error by activating reflex arcs to bring about sensory signals that match these neurons' predictions. Precision-weighting, as the determiner of where error-reducing revision happens, controls whether some error is reduced through action to alter the world and bring incoming signals into line with our prediction, or through altering our predictions to bring these into line with signals from the world.

The general view of action as "the control of perception" is pre-dated in Powers's (1973) perceptual control theory, and before that in ideomotor theories of action (Lotze, 1852; James, 1890). What is novel about PP, Clark (2013) argues, is the integration of this account of action with a theory of learning and perception, under the overarching goal of long-term prediction error minimization—though such a proposal, as noted earlier, bears interesting similarities with Hurley's (2008) shared circuits model also. Thus, as Brown et al. (2011) put it, the incorporation of action generalizes the PP scheme

> and proposes that exactly the same recursive message-passing operates in the motor system. The only difference is that prediction errors at the lowest level (in the cranial nerve nuclei and spinal cord) are also suppressed by movement, through classical reflex arcs. In this view, descending (cortico-spinal) signals are not motor commands per se but predictions of proprioceptive signals that the peripheral motor system fulfills. (2011, p. 2)

So, predictive processing is specifically the claim that perception, action, learning, and attention are implemented by the brain through predictive coding with precision weighting in a hierarchical model where predictions concern patterns over increasingly coarse spatiotemporal grain. This core is agreed on by Hohwy, Clark, and others who have developed and disputed the philosophical and cognitive scientific implications of such a model of brain functioning (Venter, 2021; Vázquez, 2020; Downey, 2018; Seth, 2014).

Where they disagree is on exactly what function this predictive processing hierarchy serves and, specifically, on the relative priority given to our two possible error-minimizing strategies of perception or action. For Hohwy, action is placed in the service of uncovering the evidence needed for more accurate perception. For Clark, perception is useful insofar as it serves the ultimate goal of successful coordination of action. In this respect, as we will see, each represents the continuation of a different tradition regarding the nature of cognition. Hohwy, as a representative of the "reconstructivist" branch, seeks to understand how the brain infers distal causal structure from impoverished sensory information. Clark, in the cybernetic and, loosely, enactive tradition cares less about such reconstruction relative to PP's utility as an explanation of how we can learn to coordinate our actions over multiple timescales.

2.2 Reconstructivist Predictive Processing

As described, early papers on predictive processing, such as Rao and Ballard (1999) and Friston (2005), focused on the model's empirical validity and efficient coding motivations. But it takes more than the unification of some physiological effects under a biologically plausible data compression strategy to get a philosopher out of bed in the morning. The aspects of PP that have arguably drawn the most attention beyond neuroscience are instead the potential epistemological consequences suggested by Friston's (2005) claim that it provides an implementation of Bayesian inference, a proposal typically linked to German physiologist Herman von Helmholtz's (1962/1866) theory of the brain as an engine of "unconscious inference."

Helmholtz saw his work as providing validation for Kant's constructivist account of experience, via discussion of optical principles that reveal the underdetermination of a perceptual experience by sensory stimulation alone. For instance, in figure 2.1, we immediately see the left-hand side as convex and the right as concave, even though the image alone is ambiguous. This

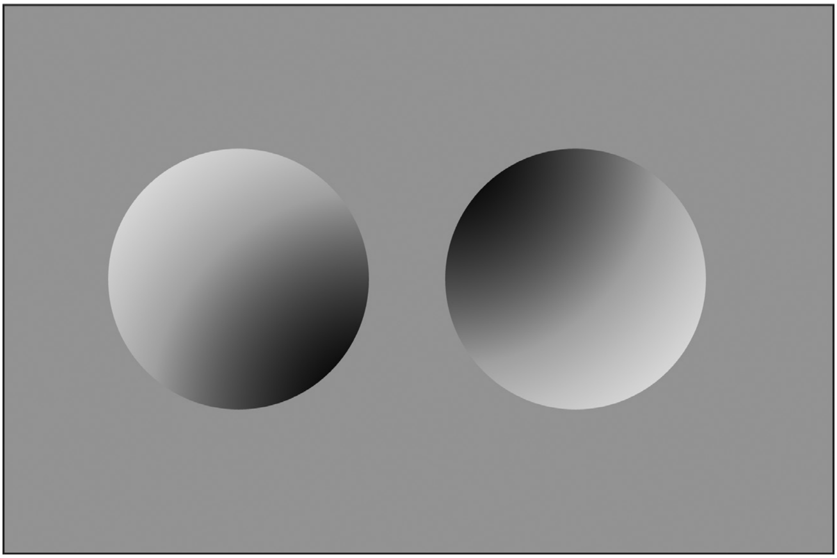


Figure 2.1
An illustration of the “light from above” prior illusion of concavity or convexity.

perceptual judgement is thus argued to be dependent on the unconscious workings of an implicit assumption that light comes from above. Another example is our ability to unconsciously discount the variability of the projection of an object onto our retina, thanks to changes in illumination and distance, in order to continuously view it as being a fixed size and color.

Being by its nature *unconscious*, our reliance on these implicit background beliefs or priors is easy to ignore until we’re confronted by an instance in which they go wrong, or conflict with other people’s priors, as in many illusions such as the famous black and blue or gold and white dress.

This idea of the brain as an unconscious inference engine came to inspire Bayesian models in perceptual psychology (Gregory, 1980), visual processing (Lee & Mumford, 2003; for reviews, see Yuille & Kersten, 2006; Rescorla, 2015), rational constructivist accounts of learning and development (Lake et al. 2017; Gopnik, 2012; Spelke & Kinzler, 2009; Tenenbaum et al., 2006), and, in machine learning, the attempt at modeling more tractable approximation strategies for how this inferential process might be implementable in the brain (Dayan et al., 1995). As Dayan et al. describe their proposal of “the Helmholtz Machine”:

> Following Helmholtz, we view the human perceptual system as a statistical inference engine whose function is to infer the probable causes of sensory input. We show that a device of this kind can learn how to perform these inferences without requiring a teacher to label each sensory input vector with its underlying causes. (p. 1)

While Helmholtz's principal concern was with the securing the validity of perception in light of Kant's transcendental idealism,[1] he did not ignore action. As he describes, "We are not leaving ourselves passively open to the [sensory] impressions intruding upon us, rather we observe, that is, we bring our organs into those conditions under which the impressions can be most precisely distinguished" (Helmholtz 1867, p. 438, quoted in Hohwy, 2013). And, as Hohwy describes this analogy of action to experiment and exploration:

> Perceptual inference allows the system to minimize prediction error and thus favour one hypothesis. On the basis of this hypothesis the system can predict how the sensory input would change, were the hypothesis correct. That is, it can test the veracity of the hypothesis by testing through agency whether the input really changes in the predicted ways. The way to do this is to stop updating the hypothesis for a while, and instead wait for action to make the input to fit the hypothesis. If this fails to happen, then the system must reconsider and eventually adopt a different or revised hypothesis. (p. 79)

If perceptions are hypotheses, then it is natural to take actions as hypothesis testing. It is via action that our perceptions of the world collide with a reality that can push back, and through action that we can gather new evidence with which to update our hypotheses through Bayesian inference.

2.2.1 Predictive Processing as Unconscious Inference

The interpretation of predictive processing suggested by Hohwy (2013) is essentially a continuation of these ideas of unconscious inference, perceptions as hypotheses, and action as hypothesis testing. On this view, prediction neurons not only encode a prediction of the signal that will be received from the next level below but also represent the distal causes responsible for producing this particular pattern in the sensory stream. The key idea is that increasing temporal depth in the patterns predicted corresponds to increasing depth in a hierarchy of distal causes. For instance, by observing changing light levels, we can not only track the circadian cycle of light to dark, corresponding to the Earth's rotation, but also a second-order, slower regularity in how this first cycle lengthens and shrinks over the course of a year. In tracking this second pattern, so the story goes, we latch on to another distal cause behind our sensory stimulation in how the position of the earth relative to the sun changes in an annual cycle. As Hohwy (2013) puts it,

> Regularities can be ordered hierarchically, from faster to slower. Levels in the hierarchy can be connected such that certain slow regularities, at higher levels, pertain to relevant lower level, faster regularities (for example, slow regularities about Aussie rules footy word frequency during the yearly news cycle pertain to faster regularities about the words I end up reading; if I know the slower regularity then I am less surprised by the occurrence of those words). A complete such hierarchy would reveal the causal structure and depth of the world—the way causes interact and nest with each other across spatiotemporal scales. (p. 28)

Thus, by hierarchical prediction error minimization, the brain not only latches onto regularities over multiple timescales but, in doing so, comes to encode a model of the hierarchical causal structure of our distal environment. According to this reconstructivist PP account (RPP), it is the rich structural content of this internal model, not the comparatively impoverished data presently streaming through the retina, that directly determines my perceptual experience. The latter is demoted to the role of model constraint, suggesting the description of perceptual experience as a process of "controlled hallucination." Among other things, such an account is proposed to explain why, when I look at the building site across the road, my experience is not a two-dimensional array of rectangular slices and flattened figures—despite this being all the information that my retina is receiving. The gray squares and silver lines match the prediction of a large three-dimensional building, and so it is this building model, not the retinal activity constraining it, that I experience.

What guarantee do we have that this prediction-error-based control is robust enough to bring our internal model meaningfully in line with reality? So far, we've only talked of prediction error minimization, but in PP these error signals are not transmitted in raw form, but always with a "precision-weighting" reflecting their estimated reliability, which directs where the error-squashing adjustment takes place. It is this precision-weighing that determines whether errors are accommodated at lower levels—for example, being explained away as the ordinary background fluctuations caused by the noise of a crackly radio signal—or whether they force deeper adjustments to long-term regularities encoded higher in the generative model, as when a persistent error signal gathers in estimated reliability, eventually trigging the realization that the words to the first section of Carl Orff's famous cantata *O Fortuna* have nothing to do with a passionate desire for tinned fish.

It is precision-weighting that guards the predictive agent against being swept to and fro by each random fluctuation in the sensory stream, or against becoming entrenched in the determined commitment to some particular pattern. Interpreted as a probabilistic measure of the reliability of a signal, it is a key ingredient in the reconstructivist interpretation of PP, allowing the error-minimization process to be cast as one of approximate Bayesian updating, in which the reliability of prior regularities learned over the course of the agent's experiential history is weighed against estimated reliability of the current evidence (more on this in the next chapter).

Does adjusting our generative model in accordance with Bayes's rule thus guarantee eventual convergence between its structure and that of the distal environment? Even in advancing the reconstructivist view of predictive processing, Jakob Hohwy is rather dubious on this point, noting that as successful minimization of error is only achieved in relation to our preselected hypotheses, reconstructivist PP becomes

> an affirmation of simple Cartesian skepticism. Since we cannot obtain an independent view of our position in the world, we cannot exclude the skeptical hypothesis that the sensory input we receive is caused by an evil, hoaxing scientist rather than the external states of affairs we normally believe in. The Bayesian framework thus entails scepticism. (2016, p. 265)

Within this reconstructivist understanding of PP, the fact that our generative model successfully predicts current sensory input becomes evidence that our model accurately captures the structure of this sensory input's causes. This is despite the fact that alternative models might also have predicted it just as successfully. These circular patterns of evidence form what Hohwy describes as an "evidentiary boundary," a point of separation between the hypothesis-generating mechanisms and the evidence that is being explained. The boundary proposed here is the edge of the sensorium: on the inside, the skull-bound brain and, on the outside, the body and world.

There are two, separate but related, issues that proponents of 4E approaches might dispute about Hohwy's internalist characterization of PP here. There is our current question of how to characterize the relationship between the systems on either side of the boundary, and there is also the question of where the boundary between a PP system and its environment is fixed—if it is indeed fixed at all. This second question is an issue for the extended mind theorist, not for the enactivist, whose more immediate concern is with the relation between mind and world, rather than with where we draw the partition between them.

As such, rejecting Hohwy's RPP in favor of an embodied, or enactivist, account, does not, as Bruineberg et al. (2018) emphasize, require rejecting that there is some meaningful boundary between a PP system and its environment, nor that this boundary *may* be drawn at the sensorimotor interface. Boundaries have, after all, been at the core of enactivist definitions of cognition from the very beginning (Varela et al., 1991). Rather, as we will see in the next section, the question is whether said boundary is best characterized as in terms of evidentiary seclusion, or of ongoing coupling, between the internal dynamics of the PP system and its environment.

Hohwy does not always keep these questions distinct, however, and thus mischaracterizes the response from proponents of enactive cognitive science to this dilemma as the suggestion that incorporating "world-engaging" action might rescue us from scepticism by breaking us out of an evidentiary boundary in order to gain "direct" access to the distal environment—a proposal he rejects. In his reconstructive story, action is placed firmly in the service of perception: a process of hypothesis-testing that allows us to intervene to control relevant variables, to seek out further evidence to resolve uncertainty, and to confirm or disconfirm our current model. Such actions (as ecological psychologists are fond of pointing out) do indeed boost our epistemic resources enough to resolve the kind of local ambiguities engineerable with the artificial constraints of the 2D images used in psychophysics labs (Orlandi, 2014). But Hohwy is correct that they could not free an RPP agent from the global underdetermination of sceptical scenarios. As he puts it:

> An appeal to action, on the prediction error scheme, reduces to an appeal to inferences about different kinds of patterns of sensory input. If a mad scientist was a hidden common cause of all that sensory input we would have no way of knowing unless she made an independent causal contribution to sensory input. (Hohwy, 2013, p. 220)

Indeed, as I'll describe in the next chapter, when framed in terms of the broader framework of free energy minimization, we see that rather than helping to pull ourselves out of the sceptical pit, the possibility of acting to alter our evidence stream only digs us in even deeper. Before we get to this, however, it's worth asking why we ever decided to jump in there in the first place.

2.3 Sensorimotor Predictive Processing

If predictive processing is supposed to be an account of how we develop an accurate representation of our distal environment, then, as we have seen, it is not a particularly reassuring one. While this interpretation of predictive processing as encoding a causal model of the external environment is common in the non-philosophical discussion also (e.g., Kanai et al. 2015), none of the central components of the PP model, described in section 2.1, entail such a view. As Orlandi (2018) argues, all we actually have is a hierarchical structure with relations of inhibition between neurons, where this inhibition either cancels out an incoming signal, or it does not. In the latter case, that signal propagates up to cause a change in a "higher-level" neuron.

So where does the representationalist reading come from? One source, as we've just seen, is a background commitment to the view of knowledge and cognition as the formation of some internal mental state that corresponds with an independent external one. Another, as Anderson and Chemero (2013) argue, may be the common fallacy of deriving semantic conclusions from merely correlational properties. There are as, they note, two senses in which we can talk of "prediction" at play in the discussion of predictive processing:

> The first sense of "prediction" (henceforth prediction1) is closely allied with the notion of correlation, as when we commonly say that the value of one variable "predicts" another (height predicts weight; education predicts income, etc.). Prediction1 is essentially model-free, and comes down to simple relationships between numbers. The second sense of "prediction" (prediction2), in contrast, is allied instead with abductive inference and hypothesis testing. Prediction2 involves such cognitively sophisticated moves as inferring the (hidden) causes of our current observations, and using that hypothesis to predict future observations, both as we passively monitor and actively intervene in the world. It is theory laden and model-rich. (p. 24)

In minimal predictive processing, the only "predictive" relationship we have is the first of these—the tendency of prediction neurons to correlate with incoming signals, in virtue of which we interpret the difference between them as an "error-signal" that is being reduced. This is the same sense of prediction by which lightning predicts thunder or the position of one coupled pendulum predicts the position of another. Such relationships of covariation may be useful to someone trying to infer the state or structure of a hidden process, but they are not inferential in themselves.

Predictive processing in itself is only a matter of developing correlations between neural activity and patterns of stimulation across a hierarchy of different temporal scales. The reconstructivist interpretation of this is extrinsic to the PP architecture and is motivated by a prior belief that reconstruction is what we are after. We need not be disappointed then by the inability of PP to deliver a guaranteed reconstruction, for we are not obliged to seek one in the first place.

As such, rather than taking action as a solution to RPP's sceptical challenge, Clark (2015) dismisses such a challenge altogether. Following embodied and enactive approaches, he suggests that the solution to the problem of perception does not lie solely in the supplementation of our inferential resources with action, but crucially in rejecting the very characterization of our perceptual goals as reconstructive in the first place. As he cites Varela, Thompson, and Rosch on this point:

> The overall concern of an enactive approach to perception is not to determine how some perceiver-independent world is to be recovered; it is, rather, to determine the common principles or lawful linkages between sensory and motor systems that explain how action can be perceptually-guided in a perceiver-dependent world. (Varela et al., 1991, p. 173)

If this is our goal, then why *should* the PP system have to infer the hidden causal nexus beyond the sensorimotor interface? To take a well-worn example, Chapman (1968) shows that a baseballer outfielder need not first model the entire onward trajectory of the baseball relative to their position, and to the field, in order to then begin the act of moving to catch it. All that is needed is the ongoing coordination strategy of "Optical Acceleration Cancellation"—that is, moving such that the ball stays at a stable position in the retinal field until it is close enough to catch. An outfielder requires no internal physics engine to recruit this strategy, no knowledge of the aerodynamic equations governing the flight of round, slightly irregular projectile in a mild northwesterly wind. All they require is an understanding of the lawlike relations between their motor output and the position of the projection on their retina.

In terms of sensorimotor PP, Clark (2015) explains, this becomes a matter of assigning high-precision weighting to errors related to the prediction that the optical projection of the ball remains at a stable location in the visual field. In such a way, the rest of the system's actions are recruited around the quashing of this particular error signal, to the neglect of most else happening on the field, until the desired state of catching the ball (or the undesired state of colliding with a teammate deploying the same strategy) is reached. Here there is no prior process of tinkering at the generative model until a lack of overall error provides adequate comfort that we've formed an accurate representation of the external world and action may now begin. Rather successful action is itself the ongoing control of a small portion of the sensory flux within those constraints that the system predicts will lead toward its target state. When understood in such a way, as Seth (2015) suggests, we can interpret this "nonreconstructivist" approach to PP as offering a mechanistic rendition of earlier embodied sensorimotor theories of perception (O'Regan & Noë, 2001).

Such "fast and frugal" strategies are much more suited toward the ongoing guidance of an organism that must constantly keep afloat in a fast-changing environment. They also fit smoothly within the rules of Bayesian optimality. As Fitzgerald et al. (2014) note, the ideal Bayesian system seeks not only to maximize empirical adequacy but also to minimize the complexity of the models recruited to do so.

For Clark (2015) the availability of locally effective nonrepresentational strategies is not an argument that we should abandon all talk of models and representations, however. Rather, the strength of sensorimotor PP is the offer of "a systematic way of combining deep, model-based flexibility with the use of multiple, fast, efficient, environmentally exploitative, routes to action and response" (2015, p. 18). For a PP system to effectively deploy such "fast and frugal" strategies as OAC, it must also be able to monitor slower-changing contextual factors (such as whether one is actually engaged in a game of baseball, or merely a participant) in order to ascertain when the circumstances are ripe for their deployment. This is why the PP system requires hierarchical depth, such that high-level states may target these large-scale increasingly invariant patterns throughout the fast fluctuations of the sensory stream.

Unlike on the reconstructivist view of PP, these high-level action-oriented representations do not allow us to "throw away the world" when we engage in planning our next action but rather to coordinate our interactions with the world over multiple timescales. We track not agent-neutral causes but agent-relative affordances at nested levels of spatiotemporal grain, from the affordance of "playing baseball" to that of "catching this particular ball."

Unlike in reconstructivist PP, securing the correctness of these "action-oriented representations" does not depend on the ability to reject the sceptical hypotheses. If a current affordance for ball-catching action is correctly detected, then deploying OAC will guide the evolution of the skilled outfielder's sensorimotor interactions to the target ball-in-hand state. This model of current sensorimotor contingencies will be successful irrespective of whether the hidden causes interacting with our sensorimotor array are instantiated by mischievous demons, curious scientists, or strange and charming fundamental particles.

2.4 What's the Point of Predictive Processing?

If the reconstructivist version of PP burdened the generative model with extravagant commitments, we may now be concerned that the action-oriented spin has been overly economical in response. Once PP is freed from the imperative of reconstruction, we still require an alternative motivation for its operations. Action is not an end in itself. To say that our predictive models are "action-oriented" and to attempt to explain perceptual contents in terms of the intention of these actions, as Hurley (1998, chapter 6) argues, merely moves the problem back a step. So, what determines the appropriateness of an action and its criteria of success?

One option is to propose that this need not be dealt with by the PP system itself, which merely describes the mechanism for achieving some function, not the function itself. That normative issue could be delegated to some separate "desire module" responsible for the calculation of an agent's goals and intentions. These can then be simply fed in as priors to a PP system tasked with bringing about their execution.

We could propose this, but to do so would undermine the entire point of the PP framework as an explanation, not just of the application of predictive models but also of their ongoing development. While PP may not itself provide us with a story about the "first priors" by which the predictive process gets started, its central explanatory payoff is as an account of the ongoing modification of these constraints through the minimization of error generated by interactions at the sensorimotor interface. If desires and intended action policies make up our prior predictions, then their selection and satisfaction conditions must be intertwined with the overall predictive economy. Thus Clark (2013) speaks approvingly of the suggestion that

> generally, personal and hedonic value is not simply a kind of add-on, implemented by what Gershman and Daw (2012, p. 296) describe as a "segregated representation of probability and utility in the brain." Instead, it seems likely that we represent the very events over which probabilities become defined in ways that ultimately fold in their personal, affective, and hedonic significance. (2013, p. 200)

While Clark (both here, and in Clark, 2019, and Nave et al., 2022) endorses this blurring together of the affective and the cognitive, he does not offer a positive story about the origin of this normativity within a PP agent. As such, and despite his sympathies with the enactive approach, the action-oriented interpretation of PP developed in Clark (2015) meets the criteria of being enactivist only in the weaker sense that the term has been applied to a variety of positions concerned with accounting for our mental life in terms of extended patterns of body-world interaction, rather than skull-bounded symbol manipulation.

While this broader use of the "enactivist" label is helpful in grouping together the shared orientation of a diversity of work in visual perception (O'Regan and Noë, 2001), anti-representationalism (Hutto and Myin, 2012), and emotion (Colombetti, 2014), it can also lead to the conflation of these narrower, cognitive scientific efforts with the more metaphysically oriented "enactive approach," first introduced in Varela, Thompson, and Rosch's (1991) *The Embodied Mind*, which a proposes thoroughgoing revision to our understanding of the mind-world relationship. This revision places a naturalized account of the emergence of normativity at the foreground of the cognitive scientist's explanatory task.

As discussed in the previous section, the central motive of the enactive approach's revisionist metaphysics concerns the replacement of the recapitulationist understanding of intentionality and meaning with a teleological one. For the enactivist, the intentionality that is the "mark of the metal" is understood not in terms of a relationship between a representational vehicle and the object it is "about," but rather, in a sense much closer to its meaning in the phenomenological tradition, as the directedness of an action toward the satisfaction of some goal.

The bioenactivist method for naturalizing this normative assessment of our actions was to seek a grounding for this normativity in the biological processes of life, then to argue that the same logic scales up to the level of the cognitive. That, as the twentieth-century phenomenologist and progenitor of the enactivist approach Hans Jonas put it, "The organic even in its lowest forms prefigures mind, and . . . mind even on its highest reaches remains part of the organic" (1966, P1).

In the next section, I will zoom out from predictive processing specifically to introduce the free energy principle (FEP) and the associated modeling framework of active inference. As a principle, the FEP proposes to "unify all adaptive autopoietic and self-organizing behaviour under one simple imperative; avoid surprises and you will last longer" (Friston, 2012, p. 2). In less cognitively laden terms, this principle amounts to the claim that a system's survival depends on it stably remaining within the same small region of possible states, and not being in states that it hasn't frequently found itself in before.

Predictive processing then stands as one possible architecture that could implement this free energy minimization process over multiple timescales. Thus, if the imperative to minimize surprises does indeed capture the kind of intentionality that bioenactivism ascribes to the autonomous organism, then the PP theorist would gain a grounding for attributing norms of successful action to the predictive brain, which it regulates its activities with respect to. In turn, PP can provide the bioenactivist with a means to scale up the basic intentionality of an autonomous organism, exemplified in behaviors such as bacterial chemotaxis, up to the rich counterfactual structure of human cognition and consciousness.

Unfortunately, this will not work. The avoidance of infrequent events falls far short of any notion of autonomy robust enough to ground the attribution of intentionality and teleological orientation to a system. Even more unfortunately, the unusually sprawling and heterogeneous structure of the free energy framework can, at times, seem constructed to hide this fact. It will take some time to see exactly why its formulation of autonomy fails. Bear with me.

3 The Free Energy Framework

Since its initial statements in Friston (2003, 2005) and Friston et al. (2006), the free energy framework has developed and mutated across thousands of publications under the hand of a diverse menagerie of coauthors. In 2022, just under two thousand papers were published on the topic, and this figure has been doubling every few years (Millidge et al., 2021a). As a result, there is now a range of subtly different, and sometimes directly contradictory, reformulations of the theory's core claims. While this rapid rate of development may be impressive, it makes any attempt at critical assessment start to look like a rather daunting prospect—all the more so given the significant time that will elapse between the completion of this manuscript and its eventual publication.

Amid this rapidly evolving literature, two independent components can be distinguished: an account of perception and action in terms of variational Bayesian inference and a "first principles" analysis of life as the homeostatic process of avoiding unlikely events—together with some connective tissue tying the execution of the former to the achievement of the latter. The common use of the label of "the free energy principle" (FEP) to refer to this entire literature is thus something of a misnomer. Henceforth, I will restrict the label of "principle" to refer to the second component alone, as distinguished from the model of perception and action that is commonly referred to separately as "active inference," and I will refer to the package of both together as "the free energy framework"

It is the free energy *principle*'s (FEP's) account of life as homeostasis that underpins Friston and colleagues' claim to have identified a fundamental theory from which bioenactivist accounts of organismic self-production can be derived (Allen & Friston, 2018; Ramstead et al. 2021)—a principle so basic that as Friston et al. (2012) proposes, "there is no need to recourse to any other principles" (p. 2). It is this view of organisms as essentially self-stabilizing systems that will form the main critical target of this book in chapter 6 onward.

While this conceptual analysis of biological survival is distinct from the modeling framework of active inference, the FEP's connection to this account of how perception, action, and inference are entangled and predictively coordinated across multiple timescales is often taken as a central advantage that sets it apart from other accounts of biological self-organization—accounts that are commonly criticized for their inability to "scale up" to higher-order cognitive processes. As such, chapters 3 to 5 will first develop this connection to show how *if* organisms are essentially homeostatic systems, then they could be understood as entailing the kind of statistical models that allow their behavior to be described as minimizing free energy via active inference.

Before developing these connections, I should emphasize that just as the FEP is distinct from active inference, so too can the modeling tools used in active inference be developed independently of the success or failure of the FEP. Indeed, active inference is sometimes used to refer to nothing more than a set of formalisms for redescribing the dynamics of coupled systems in terms of statistical models. When framed as a philosophically neutral modeling framework, the only important concern is how well these tools can describe and predict the behavior of some system of interest—and the systems and processes to which they have been applied are broad indeed: from neural dynamics (Friston et al., 2017; Da Costa, et al., 2021) to cognitive and behavioral phenomena (Parr & Friston, 2017; Friston et al., 2016) to social coordination (Friston et al., 2020; Constant et al., 2019), self-organization (Friston, 2013, 2018), and even the climate (Rubin et al. 2020).

Advocates of active inference have not restricted themselves to such a philosophically neutral presentation, however, and the framework is best known as a specific proposal about how perception-action are actually coordinated by a cognitive agent. Predictive processing is one example of this general approach, as a proposed architecture by means of which our brain could implement the process of model-updating that active inference describes. In this context, the statistical descriptions of active inference are no longer presented as merely being a useful model that can be employed by us as observers of a system (Bruineberg et al., 2022; Andrews, 2021; van Es & Hipólito, 2020). Instead, the dynamics from which said model was derived are interpreted as actually encoding said model as a "belief" that it uses it to perform inference or direct actions (Kiefer & Hohwy, 2019; Ramstead et al., 2020). It is this realist interpretation that turns the mere modeling framework into a "unified theory of cognition" as a process of approximate Bayesian inference—specifically variational inference (Buckley, 2017).[1]

One way into this "unified theory," referred to by Parr et al. (2022) as the "low road" to active inference, simply begins with the assumption that the problems of cognition, perception, action, planning, and so forth are all problems of inference. Given that, in many cases, exact Bayesian inference will be intractable, active inference is motivated as an account of how a system can approximate this solution via the minimization of the statistical quantity of free energy. So, the justification for describing the system as an approximator to Bayesian inference is based on the fact that it cannot actually be an optimizer, but the idea that it needs to solve an inference problem at all is the starting presupposition, rather than the conclusion, of this "low road" approach to active inference.

This is not an unreasonable starting place. The presumption that cognition, perception, and action are all processes of inference is common beyond the free energy and active inference literature, as a shared commitment of "Bayesian brain" approaches. This is, however, an assumption that runs into conflict with the enactive approach to cognition. As I argued in chapter 1, enactivism should not necessarily be understood as forbidding the attribution of representations, internal models, or inferential processes to a system but rather as the attempt to set criteria for when, if ever, we are justified in attributing any purpose or a function to a system—inferential or otherwise. From this standpoint, the fact that some behavior can be described *as if* it is solving an inference problem is not enough to justify the claim that the behaving system is indeed genuinely trying to solve that problem.

Rather than ignoring this challenge, as the "low road" approach of the Bayesian brain does, the "high road" approach to active inference draws on the free energy principle to tackle the enactivist challenge head on. If the FEP could directly establish that free energy minimization is something that every living system must do to survive, then, so the argument goes, this would provide justification for attributing this approximation process, together with the models it implies, to living systems themselves.

As Ramstead et al. (2018) state:

> The "intentionality" or "aboutness" of living systems—that is, the directedness of the organism towards a meaningful world of significance and valence—emerges as a natural consequence of embedded adaptive systems that satisfy the constraints of the free energy formulation. For a living thing to be intentional just means that it entails a generative model . . . Put simply: active systems are alive if, and only if, there [sic] active inference entails a generative model. This makes the generative model of central importance to the free energy formulation, since it defines the form of life that an organism is seen to enact. (Ramstead et al., 2018, sup. mat. 4, p. 33)

And Wiese and Friston (2021) express a similar point, claiming that:

> In other words, the FEP provides a (very general) answer to the question "What *does* a system do, when it stays alive?" as well as to the question "What *should* a system do, in order to stay alive?" The FEP thereby accounts for a basic form of goal-directedness that is characteristic for basic forms of intentionality. (p. 7)

Without this principled claim about the connection between inference and survival (or at least some other justification for attributing this function to the system) the description of said system as a predictive modeler or free energy-minimizer would be supported by nothing more than a formal similarity between its stable dynamics and the computational process of approximate Bayesian inference. As will be familiar from triviality arguments against the computational theory of mind, such structural similarities are worth little on their own and are insufficient to attribute internal models and anticipatory contents (Sprevak, 2018).

If the FEP *does* adequately formalize the intentionality of living systems, however, this would not only allow active inference to be proposed as a theory of what our brains (and bodies) are actually doing. In addition, it would position the theory to aid the bioenactivist in turn, by connecting their account of basic intentionality up to a theory of sensorimotor learning and coordination that might support the formulation of enactive theories of higher cognitive processes.

In this chapter, I will describe the basic components of how active inference redescribes perception and action as a process of statistical inference and reduces this to the maintenance of stability in the face of perturbation. In the next chapter, I will describe the justification for the free energy principle's claim that this maintenance of stability is the defining feature of living systems, and in the following chapter, I will explain more recent attempts to extend active inference beyond the brain, by using the idea of conditional independence between internal and external states to formalize the notion of a sensorimotor boundary. This gives us a two-part definition of a system in terms of two forms of statistical stability: (1) the stability of a probability density over most likely states of the various parts of a system and (2) the stability of interactions between these parts that preserves a statistical boundary between the system and its environment.

All of this sets us up for the second half of this book, in which I will analyze the prospects of Friston's (2019b) "existential dyad" against two yardsticks. The first is how well these criteria allow us to capture the bioenactivist concept of autonomy. The second is how well they capture the essential features of living systems, independently of any prior commitment to the bioenactivist picture. I will argue that it fails at both. Such requirements are both too general to capture the specific features that imbue a living system's activities with an intentional orientation and, at the same time, too restrictive to shadow the amorphous organism in its improvisational dance between the formal and material constraints that shackle inanimate substances.

3.1 Variational Inference

The first component of the inferentialist part of the free energy story is a development of variational inference[2]—a strategy taken from machine learning for approximating a solution to an intractable inference problem (Hinton & van Camp, 1993; Neal & Hinton, 1998, and see Beal, 2003 for a more recent development and overview).

In the case of the brain, we might think of this task in terms of finding the probability distribution over hidden causes that best predicts our sensory observations.[3] The optimal way to do this would be to begin with a joint distribution [$\mathrm{P}(O,H)$] over the probability of each observation [O] and each possible value of the hidden variable(s) [H]. To calculate the probability of a specific hypothesis [H_1] given a specific observation [O_1], we need to decompose this joint distribution into the prior [$\mathrm{P}(H_1)$], the likelihood [$\mathrm{P}(O_1|H_1)$], and marginal [$\mathrm{P}(O_1)$].

$$P(H_1|O_1) = P(O_1|H_1)P(H_1)/P(O_1) \tag{1}$$

Doing this for each hypothesis [H_i], we can then use this formula to update our probability distribution over hidden causes on the basis of each new observation [O_i], to get a new posterior [$\mathrm{P}(H|O_i)$], which we then feed in as our new prior [$\mathrm{P}(H)$] for the next round of observations [Oi+1] and updates.

Eventually, all being well, this process will stabilize, the posterior given $Oi + \mathrm{n}$ will be the same as the prior from the previous update, and we can take ourselves to have inferred the hidden causes behind our observations.

The problem with this process is that calculating the marginal $[\mathrm{P}(O_1)]$ requires integrating across all the different possible hypotheses $[\mathrm{P}(O_1,H_1)$, $\mathrm{P}(O_1,H_2)$, . . . $\mathrm{P}(O_1,H_n)]$ in order to figure out how likely some particular observation is “in general.” Such an operation quickly becomes intractable with anything like the number of possible hypotheses the brain has to deal with.

So, rather than attempting to infer the true posterior distribution from the entire space of possibility distributions, an alternative strategy is to begin with a restricted set of simple distributions that can be characterized by a small number of parameters—for instance, a Gaussian distribution that can be encoded with just the parameters of mean and variance. The selected distribution is termed “the recognition model,” or “variational model”[4] and its parameters can then be progressively tweaked to perform gradient descent/ascent over two elements. The first is typically referred to as “accuracy,” but, as I will explain, this label is misleading and I will call it “predictive adequacy,” or adequacy for short.

This quantifies the likelihood of our evidence under the recognition model, which is what we want to maximize by changing the parameters of said model. The second, *complexity*, quantifies the amount that we are changing the recognition model, which we want to minimize. The name “complexity” reflects the fact that if we do not attempt to limit how much we tweak our model in order to best fit each new observation, then we may end up overcomplicating it to the point that it is too specific to capture general regularities in our observations.

These two quantities thus reflect what is called the bias-variance trade-off (Geman et al., 1992), between (over)fitting our model to each new variation versus (over)generalizing from prior regularities. Balancing these quantities prevents us from reassigning all our confidence on the single hypothesis that makes our current observation most likely, which would maximize adequacy, or digging in our heels with the model we have already learned from previous observations to minimize complexity. These are exactly the two considerations that are weighed against each other under true Bayesian inference. By attempting to minimize these with respect to a selected distribution, we have thus converted our intractable inference problem to a much simpler optimization one.

The rather powerful sounding “free energy” (Friston, 2005, 2010) that gives the FEP its name is simply an alternative label for the combination of these two functions.[5] When encountered in machine learning or statistics, it would more likely go by the name of the (negative) “Evidence Lower Bound (ELBO)” It can be written, with tolerance for a little simplification, like this:

$$F = D(q(H) \mid\mid p(H)) — < \ln p(O|H) > q \tag{2}$$

The first term, $\mathrm{D}(\mathrm{q}(H)||\mathrm{p}(H))$, is the divergence between the recognition distribution [$\mathrm{q}(H)$] and our prior distribution [$\mathrm{p}(H)$] over hidden states. Specifically, this is quantified by a metric called Kullback-Leibler or KL-divergence. This is the “complexity” described earlier, representing how significant a change we are making to our model in order to reduce free energy with respect to this new observation. The second term, $<\ln \mathrm{p}(O|H)>q$, is the “accuracy” or “adequacy”—which captures how probable some recognition distribution over hypotheses makes our observations.[6]

This has been converted to a logarithmic scale, so, rather than ranging from 0 to 1, it takes a negative value up to a maximum of 0, which represents certainty. The more unlikely the observation the lower the log probability—that is, the further this negative value is from zero. Because subtracting a negative number is equivalent to adding a positive one, the lower the adequacy the greater the amount we are adding to our free energy equation.

Hence, by minimizing complexity and maximizing “adequacy,” free energy is reduced.

I want to stress that while the second quantity is typically labeled “accuracy,” this is somewhat misleading here, for this captures neither how close our recognition model is to the generative model that we’re aiming to approximate nor the true statistics of the process generating our observations. All it measures is how successful our recognition model is at predicting observations. Like a broken clock or a lucky beginner, a model may have predictive adequacy for some particular observation while still being divergent from the true statistical properties of the process that produced that observation. With enough observations, however, the hope is that persistently high predictive adequacy serves as evidence for the fit between our model and the hidden variables that cause our observations.

What’s neat about free energy is that it can also be broken down differently into what I will call (1) *divergence*: the difference between the recognition density and the exact posterior of true Bayesian inference, and

(2) *surprisal*: the unlikeliness of some sensory data relative to the actual generative model—the model that we are aiming to tweak our recognition model to more closely resemble.[7] This looks (again with a little simplification) like this:

$$F = D(q(H)\ ||\ p(H|o)) + (-\ln p(o)) \tag{3}$$

There are a few points worth highlighting here. The first point is that while, like in equation 2, the first term of equation 3 is also a KL-divergence metric; in this case, it is between the recognition model and the exact posterior of true Bayesian inference [$p(H|o)$]. In equation 2, in contrast, it was the divergence between the recognition model and the prior [$p(H)$].

The second point to highlight is that we are adding surprisal as a negative log probability [$- \ln p(o)$] rather than subtracting a log probability directly. This fussing about with signs looks a bit strange, but all it is doing is converting a maximization requirement into an equivalent minimization one—a standard practice when dealing with optimization problems. To understand this, you just need to remember that log probabilities are negative numbers, which are maximized at 0 for a certain event. Multiplying log probabilities by -1 merely flips this sign around so that the values of this surprisal term are now positive numbers that are minimized at 0 instead.

In either case our free energy functional will be minimized when this term is closest to zero. In the case of a log probability, the lower this negative value is, the further we are from zero, and so we want to maximize it. In the case of the positive values of the negative log probability that makes up surprisal, lower numbers will be closer to zero, and so we want to minimize this quantity. Thus, the advantage of formulating free energy in terms of the addition of surprisal is that it means that both terms of our minimization task are themselves minimization problems.

A second, and more important, point is that, while adequacy and surprisal seem similar, they are not just the inverse of each other. Adequacy here is a function of an observation and our recognition model—our current best guess at the approximate statistics of the underlying process that generated this observation. It is something the free energy minimizer is able to measure and reduce. In contrast, surprisal (as it occurs in the FEP) is improbability under the exact posterior of the full generative model—which is what we are aiming to approximate. High adequacy entails low surprisal only when our approximate recognition model is sufficiently close to the generative model.

Somewhat confusingly then, it is what has been termed "adequacy" here, not surprisal, that best corresponds to what we would ordinarily call the personal-level "suprisingness" of some particular piece of evidence in relation to some inferring agent's expectations. The technical notion of the *surprisal* of some observation is inaccessible to the agent for two reasons. The first, already discussed, is because it depends on an intractable marginalization of the generative model P (O, H) to give the "overall" probability of $P(O_1)$ under this model. The whole point of variational inference is to avoid this marginalization.

A second reason is that, as we will see in the next second, on some presentations of the FEP, the generative model is not actually something encoded by the agent itself at all.

So, while the division into "divergence + surprisal" of equation 3 is not accessible to the free energy minimizing brain, what dividing things up this way shows *us* is that by minimizing free energy the brain can limit the possible divergence between its recognition density and the true posterior—to the extent that, if the latter were in the restricted class of simple distributions, they will become equivalent. When there is no divergence between our recognition density and the true posterior, then free energy reduces to surprisal. This allows surprisal minimization with a true model to be cast as a special case of free energy minimization, a point that will become important when we consider the generalization of the FEP beyond the brain. As it is, however, all the free energy minimizer can rely on is that minimizing free energy will push its recognition density toward the closest thing it can get to the true posterior from within a constrained class of simpler, tractable distributions.

How does all this connect to predictive processing? Well, the variational inference procedure does not specify which constraints we should select for our restricted class of probability distributions. But, if we choose the recognition distribution appropriately, then said procedure may be implemented by a predictive processer. The first assumption we need is the *Laplace assumption*, which restricts our class of possible recognition distributions to Gaussians, which can be parameterized with just the sufficient statistics of mean and variance. The second is the *mean-field approximation*, which assumes that the posterior distribution across all the states of all hidden variables can be decomposed into a number of separate distributions across the state of each independently. With these choices, each prediction in PP can be interpreted as the mean of one of these distributions, the precision as encoding its variance, and the overall precision-weighted prediction error of the input as the free energy for our present evidence relative to said distribution (Gershman, 2019).

3.2 Incorporating Action

Through variational inference, we can tease out the assumptions and constraints that motivate the description of predictive processing as an approximation of Bayesian inference. This is nice, but if this were all there is to the free energy framework, it would not be all that novel—the suggestion that variational methods could explicate how the brain performs "unconscious inference" dates back to 1995 and Dayan, Hinton, and Abbott's proposal of the "Helmholtz Machine." Nor is it what we went looking for when we turned to the free energy principle in the attempt to provide an alternative, nonreconstructive, imperative for the operations of a PP system.

What differentiates active inference, unsurprisingly, is the incorporation of action and the attempt to describe an agent that can interfere with its sensory input, in contrast to the standard variational inferrer who can only passively track whatever regularities it is presented with. To see why action is important, we can note that variational inference describes only how a system can update its recognition model to more closely resemble a joint probability distribution over observations and hidden causes—the generative model. This kind of probability distribution over states is the only thing that an active inferrer can represent or approximate. We may adjust this model to better capture the regularities in the generative process that produced our observations, but for this to work there must *be* such regularities. In other words, the statistical behavior of the generative process must remain stable if we are to succeed in inferring it.

The environment around us does not seem to provide any reason to embrace such a universal constraint. There are oscillations and orbits, but there is also growth, development, and collapse—recurrence not guaranteed. In the grand thermodynamic stream of things, the stability we observe is but a temporary eddy in the overarching flow toward disintegration. If it makes sense to model our environment, one in which systems are stable and surprisal is minimized, then it seems we need to look first at how and why we act to make our little corner of the world this way. To see how active inference answers this, we'll now return to the debate of whether the predictive brain prioritizes representation or action-guidance, now aided with the various decompositions of free energy.

3.2.1 Generative Cycles and Active Systems

Suppose you had heard Edinburgh often referred to as "the Athens of the North,"[8] and, mistaking this to be a matter of climate rather than an unedifying comparison between the unfinished nineteenth-century folly of Calton Hill and the two-millennia-old Greek Acropolis, you moved there as an escape from the Manchester drizzle.[9] Arriving in the city in full expectation of balmy 21°C sunshine, you find yourself confronting the persistent error of a dreich sub-zero day.

You could just accept your fate and update your model to incorporate an increased probability of states at the lower end of the temperature scale. Alternatively, you could reduce this (rather embarrassing) error by heading to the airport and flying to warmer climes. Both would serve to minimize the free energy of your ongoing observations, but it is hard to make sense of why you'd actively adjust your experience in order to fit a prior hypothesis if your aim was to form an accurate model of the external environment. If this was your ultimate goal, then all your error-avoiding behavior seems to have done is prevent you from learning an important lesson about the untrustworthiness of tourist boards.

This problem is familiar from the earlier discussion of predictive processing, but things get even messier if, as suggested in the previous section, we understand the generative model not as something encoded in an inferring brain, but rather as a description of the actual statistical properties of the observation-generating process that we are trying to infer.

Suppose as you exited Waverley train station you instead found yourself captivated by Edinburgh's medieval skyline and decided to stay. Unable to tolerate the city's outdoor temperatures, you move into a flat with a castle view and a magnificent old fireplace, then install yourself comfortably beside it. Congratulations! You have minimized prediction error relative to your goal state of 21°C—but this isn't all that you've done. In lighting up the fireplace, you have not only changed your current state to bring it in line with your prediction but also altered the long-term statistics of the environment that you are attempting to model, making 21°C a more likely state to encounter in the future.

Every evening, all across the city, thousands of other residents do the same—though these days a gas boiler is the more common method. As a result, the average outside air temperature of Edinburgh city center is a few degrees higher than in the surrounding countryside—and getting warmer (Price, 1979). If our heat-seeking denizens continue to multiply in number and energy consumption, then, climate-change permitting, you may one day be able to emerge blearily from your Georgian terrace onto the streets of New Town and find your prediction of a 21°C air temperature to be perfectly satisfied. Your internal recognition model of Edinburgh's average air temperature is now accurate—but only because the fact that you assigned this temperature a high probability drove you to make the world conform with it.

Incorporating action into the free energy framework unleashes a strange circularity. A circularity that undermines the Helmholtzian understanding of the process whose statistics the generative model describes (and which the brain is supposed to partially approximate in its recognition model) as being composed solely of distal environmental causes. Allowing agents to act to change their sensory input unavoidably inserts the modeling agent into the very observation-generating process that this agent is simultaneously attempting to model.

If we are modeling anything under active inference, then it is not the agent-independent world but rather the fused self-world system. Thus, the hidden causes "H" described in our generative and recognition models now refer not only to external states of affairs (*E*) but also incorporates our own actions (*A*), actions that were determined by our internal model (*I*), which depends, in turn, on our sensory observations (*S*).

Rather than a clean divide, with the observation-generating process of hidden environmental causes on one side and an internal action-generating model on the other, what we now have is a generative cycle, from an internal "model," to actions, to external causes, to sensory observations, and then back to internal states of the agent. It is the formalization of this that makes the FEP a marked departure from the standard variational procedure—in which the inferring system has no effect on the stable statistics of the system whose behavior it is attempting to infer.

In discussing the problem of evil demon scenarios for PP, Hohwy (2016)[10] describes the circularity inherent to predictive processing, where our evidence is taken to confirm our prior hypotheses so long as it is consistent with them, irrespective of whether it would rule out other equally probable alternatives. The problem looks much worse when our evidence is not only interpreted, not only selected, but actively *created* as a result of these hypotheses. As Bruineberg et al, (2018) point out, if the free energy minimizing brain is a hypothesis-testing scientist, then it starts to look like a "crooked and fraudulent scientist" who "decides on the outcome of an experiment beforehand . . . and manipulates the experiment until the desired result is reached" (p. 2430).

How then can this crooked scientist picture be reconciled with the subsumption of perception and action under the logic of approximate Bayesian inference that is often supposed to be free energy framework's crowning glory?

Firstly, it should be noted that the kind of crookedness just described does not undermine a Bayesian interpretation of the recognition model. A good Bayesian is not accountable for the evidence she receives, only what she does with it. In consistently gravitating toward sources of warmth, you can curate an evidence stream that would direct a bona fide Bayesian process to produce a model of the average air temperature that is 21°C. Still, once there are no pre-fixed independent statistics to be inferred, and once the observations made are determined by the prior expectations of the modeling system, this inferentialist understanding looks less helpful as a way of interpreting the success of a free energy minimizing system in preserving itself.

To ameliorate the uncomfortable subjectivity of setting initial priors in Bayesian inference, it is often pointed out that, even when agents begin with wildly divergent priors, these differences can eventually get "washed out" through the process of updating on the *same* evidence, leading to convergence in their models. Agents that create and curate their own personally tailored evidence stream to support their initial priors undermine this possibility. The incorporation of action appears to have left us rather unmoored. When we can change not just our internal models at will, but also the structure of the process generating our observations, then we appear to lack any fixed restrictions on how we should proceed or any guarantee that different strategies will converge on the same results.

We can try to regain some stable footing here by noting that neither system, neither brain nor environment, is infinitely flexible. We might be able to alter our surrounding temperature to some degree, but this environmental component of our generative cycle also constrains the extent to which we can bring about observations that fit our internal expectations.

Consider the other side of the world where 5.6 million Singaporeans also struggle with their local climate—though from a more tropical standpoint than Edinburgh's inhabitants. The city-state has the among the highest number of air conditioners per capita in the world, allowing residents to shuttle between apartment buildings, office blocks and underground malls, chilled to a cool 18°C. Despite these efforts, indeed, partially because of them, Singapore is growing warmer at twice the rate of the rest of the world (Jiang et al, 2021). While the action of the Edinburgh resident in turning up the thermostat amounts to a self-reinforcing prediction, the Singaporean's air conditioning actions are unable to overcome the constraints of thermodynamics. In this positive feedback loop between climate change and air conditioner adoption, the attempt to realize a predicted goal temperature in the short term actively undermines the ability to achieve it over longer timescales (Howarth, 2020).

Bringing predicting agent and predicted world into alignment can no more result from the dogged pursual of prior expectations in the face of their repeated failure than it can from passive conformity to whatever sensory evidence our environment throws at us. You can't infer a silk purse from a sow's ear—for all that a pigskin wallet might be on the table. Still, the fact that our repertoire of prediction-fulfilling actions is limited by external constraints does not yet solve our problem. Even within these limits, the space of possibilities is still underconstrained. In most situations, the active inferrer will be presented with a choice between acting to bring its observations in line with its recognition model and changing this internal model to fit its observations. In the former mode, we may still think of it as inferring an accurate model—albeit a model that includes the likelihood and consequences of its own actions. In the latter, it looks more like a system that is attempting to sculpt the world in line with its model. How is this trade-off settled?

3.2.2 Act First, Think Later

We have two ways to minimize free energy or prediction error. We can act to change the world, and so the generative model, or we can update our internal recognition model to better predict our observations, and thus, ideally, to better resemble the generative model. In the context of PP, this is typically expressed as the claim that "perceiving the world (perceptual inference) and acting on it (active inference)[11] turn out to be two sides of the same coin." (Gładziejewski, 2016, p. 562). As far as PP is concerned, both action and perception involve changing variables in order to minimize prediction error—the only difference is that the former changes concern external variables and the latter involve only internal changes.

The free energy framework complicates this, however, by revealing prediction error/free energy to be a compositional quantity—one that may be broken down into (a) *adequacy*: how likely some change to our recognition model makes this input and (b) *complexity:* a quantification of the degree to which we're changing our model to better incorporate each new input. With this decomposition in hand, we can identify a functional asymmetry between perception and action, revealing that they are not just interchangeable means to the same end. If we choose to improve adequacy via perceptual updates to our model, then it comes with a complexity cost—we optimize one of free energy components at the expense of increasing another. If instead we just act to change our input, then we improve adequacy for free!

We can also look at this trade-off from the perspective of the alternative decomposition of free energy, into (a) *divergence*: the difference between our approximate recognition model and the ideal Bayesian's model of the causes behind our sensory input, plus (b) *surprisal*: the true unlikeliness of this particular sensory state for this system. While we cannot measure or reduce either of these directly, we do know that surprisal has nothing to do with our internal recognition model. While updating our model via perceptual inference may bring it closer to the ideal Bayesian posterior, and thus make predictive accuracy a better proxy for surprisal, it cannot directly reduce surprisal. In contrast, reducing free energy via action not only allows an agent to reduce the surprisal of its observations, as seen in the previous sections, it also inserts the agent itself into the statistical structure of the process it is trying to infer, altering the probabilities of the different observations it is likely to receive. So, by bringing the long-term statistics of the generative process closer to what the agent expects, action can *also* reduce the divergence between the correct model of this process and our simplified recognition model.

On the one hand, action improves adequacy with no complexity cost, and on the other, it can potentially reduce both surprisal *and* divergence, where perceptual model updating will reduce only the former. So, free energy minimization clearly prioritizes action over model updating. Still, as we saw in the previous section, this is a strange confirmation-seeking form of action, entirely unlike the hypothesis-testing process that Helmholtz proposed. What we have now is not just an agent that *may*, on occasion, choose to alter the world to fit its model but rather an agent for whom this is the ultimate goal. For such an agent, the only reason to update its model is as a last resort, when it finds itself unable to actively assert its expectations onto the world—as in the case of the positive feedback loop between increased air conditioning and global warming.

From a perspective that views cognition in terms of the epistemic norms of understanding and accuracy, this seems like unacceptable behavior. In the next chapter, we'll look at how such a strategy might make more sense when approached in terms of biological norms of adaptation and survival. Before we get to that, however, we need to complicate things a little further by incorporating hierarchical active inference and parameter learning, to see whether this can allow active inference to account for an agent that does more than doggedly enforce its initial expectations onto the world.

3.3 Active Inference and Learning

The discussion of active inference thus far has been limited to the process of updating a single recognition model, to bring this closer to the ideal posterior of Bayesian inference without needing to perform an intractable marginalization of a full generative model. This variational procedure avoids marginalization but does not remove the need for a generative model altogether. This joint distribution [P(O, H)] is still required to obtain the likelihood of particular observations given hidden causes (hidden causes that include the agent's own actions). This is needed to calculate free energy for a recognition model and an observation using the "accessible" rearrangement given in equation 2. We have said nothing, however about how this generative model is formed. Instead, we have simply presumed that the system is already somehow endowed with a "pre-ordained" generative model.

We've seen that what makes the internal recognition model of an active inferrer good or bad is not a question of whether it accurately describes the most likely mind-independent state of affairs, but rather whether the state of affairs it treats as most likely is one that can be easily secured by the agent's actions. When it comes to the generative model, however, there must be a degree of veridicality in how this depicts the relationships between observations and their potential causes, in order for the agent to be able to assess which actions will bring about the observations it predicts.

As such work in active inference standardly presumes not only that the agent has a generative model but that it is a "*veridical* generative model" (Friston et al., 2015; see also, Friston et al., 2017; Hesp et al., 2021; Buckley et al., 2017). As Parr et al. (2022) describe this point in their textbook on active inference:

> In many practical applications discussed in this book, we simply assume that the dynamics of the generative process are the same as, or very similar to, the generative model. In other words, we generally assume that the agent's generative model closely mimics the process that generates its observations. (p. 131)

While they qualify that their intention is not to say an agent must already have a *perfect* model in order to engage in active inference, the examples they give of inference with an "imperfect model," concern inaccuracies in the agent's initial estimations regarding the present state of the world and their place in it, while the structure and parameters of its generative model are still assumed to be veridical. As Raja et al. (2021) point out, this presumption that we already have a good generative model of the world to explain how we learn through perception, looks like a vicious circularity in the FEP's explanation. Surely the very thing we want an account of perception to explain is how this model developed in the first place.

Still, while applications of active inference may typically have focused on inference with a predetermined veridical generative model, Parr et al. (2022) claim that this assumption is merely a convenient simplification for practical purposes. Once we move to the kind of hierarchical structure we found in predictive processing, they suggest that is relatively straightforward to make sense of how a generative model can be learnt via exactly the same process that was used to optimize the recognition model in the previous chapter. What we have now is simply the same process of free energy minimization, occurring over a different temporal scale, in which our probability distribution ranges over model parameters, rather than over hidden causes directly. Or, as they nicely describe this idea:

> Inference describes (fast) changes of our beliefs about model states—for example, how we update our belief that there is an apple in front of us after observing something red. Learning describes (slow) changes of our beliefs about model parameters—for example, how we update our likelihood distribution to increase the value of the apples-red mapping after observing several occurrences of red apples. (Parr et al. 2022, p. 129)

So, active inference accounts for the learning of model parameters that encode the relationship between lower-level "state of the world" variables, via the explicit introduction of probability distributions over these parameters. These probability distributions are then updated on slower timescales but, crucially, still via the same process of variational free energy minimization. This separation of timescales between learning and inference, Parr et al. suggest, might be mapped to the separation of timescales between neural dynamics and synaptic plasticity.

As I will describe in more detail in chapter 9, this hierarchical extension introduces a greater degree of flexibility into the free energy framework by providing a means to incorporate model updating as a directed process, rather than just a last resort when an agent cannot act to make the external world fit its prior model. By incorporating a hierarchy, we can now treat the revision in the probability assignments at one level as the "action" of a higher-level model, directed toward the reduction of free energy by making this lower-level model conform with the expectations of the higher-level model.

Still, while this hierarchical extension allows us to describe how a generative model can change over time, as Rutar et al. (2022) note, it only concerns parameter revision within a fixed state space, a framework that is insufficient to account for open-ended processes like learning and development. I will return to the need to conceptualize learning and development as open-ended processes in chapters 8 and 9, where I argue that this open-endedness is the distinguishing feature of vital processes, one that any enactive theory of cognition must account for.

For now, we can temporarily allow that a fixed higher-level model of dependencies between variables with a fixed space of possible states may offer a valid description of the process that interests us. To suspend doubts about the validity of this representation does not resolve questions about how an agent could come to have knowledge of it. Hierarchical accounts of model-learning do not resolve this need for such a preordained and veridical model—they merely pass the buck on explaining where it comes from.

To define the free energy of model parameters, we still require the assumption of a further higher-level generative model over said parameters, relative to which this is defined. If we then ask where this second-level model came from, the only possible FEP-based answer is that it was also learned via free energy minimization—an answer merely moves the problem of learning back yet another step with the postulation of yet another level of generative model.

For all that the move to a hierarchical structure allows for lower-level changes in the model to be described as learning (via free energy minimization with respect to a higher-level model), no matter how many layers we introduce, there must eventually be a final level. To get started with free energy minimization and active inference, we still require a predetermined and predeveloped model that cannot itself have either been learned via free energy minimization or be amenable to subsequent change by free energy minimization.

Once we have a principled basis to establish the structure of this ultimate generative model for a particular system, then we can approach it in terms of the active inference framework. To do this, then we need to backtrack along the "high road" that connects the FEP and active inference, to see how the former aims to ground the identification of probabilistic models in the structure and dynamics of biological systems.

3.4 Where Does the First Generative Model Come From?

The question of how an agent can emerge pre-adapted into its world is far from a unique issue for the free energy principle. As the central question of modern biology, it is one we already have a well-accepted answer to in the form of genetic inheritance and natural selection. As such, the standard line in the free energy literature is that agents inherit this first model from their ancestors, for instance, by genetic inheritance (Friston et al. 2015) or development in utero (Ciaunica et al., 2021). Insofar as an ancestor would only be expected to survive if it had a good generative model, so we can expect our inherited model to be apt also. In some papers, it is suggested not only that natural selection favors the preservation of good generative models but that natural selection is itself an inferential process (Ramstead et al., 2018; Kirchhoff et al., 2018). While the idea that natural selection can be described in terms of updates to probability distributions over possible models is not a new idea, FEP advocates make a stronger claim, namely that the concept of reproductive fitness is itself derived from free energy minimization as a more foundational principle.

The connection between survival and the preservation of a generative model will be developed in the next chapter and the idea that fitness and adaptivity either derives from, is reducible to, or is describable by, probabilistic principles will be criticized in chapter 9. Putting aside how these evolutionary processes might, or might not, guarantee the “goodness” of a generative model for now, the question I want to focus on here is how a model can be the kind of thing that an agent is born with and that is inherited across generations.

Such an understanding moves us away from the account of generative models developed in predictive processing where these were understood as structural models, encoded in a specific individual’s brain by top-down synaptic connections (Kiefer and Hohwy, 2018, 2019). In contrast, Ramstead et al. (2020) assert that “under the FEP, generative models are not explicitly encoded by physical states. That is, they are not encoded by states of the brain. Rather, it is the adaptive behaviour of the system that implements or instantiates a generative model” (p. 231).

This claim evinces a rather strange metaphysics of models and raises a number of confusing issues.

The first point of confusion is that, as generative models are typically understood in predictive processing and the active inference literature, there is a clean distinction between the structure of this internally encoded model and the external hidden causes that constitute the generative process (Friston et al., 2016; Parr et al., 2022). Despite their rejection of the idea that generative models are internally encoded, Ramstead et al. (2020) continue to support this distinction, insofar as “actual causes of sensory input depend on action (i.e., on a generative process), while action depends on inference (i.e., on a generative model” (p. 234).

As described in section 3.2, the circularity of active inference—in which observations are produced by actions, which were in turn caused by the internal generative model—makes such a distinction hard to sustain for the generative model is now itself a part of the generative process that produces an agent’s sensory input. This confusion is reflected in the fact that, at some points, Ramstead et al. (2020) describe actions as part of the hidden causes that are *described by* the generative model, yet at others, as in the quote above, this adaptive behavior is instead cast as the vehicle that *instantiates* or *implements* this model.

The suggestion seems to be that we can preserve this model-process distinction in terms of the direction of dependencies, where the dependence of sensory input on action is the "process" and the dependence of action on internal states is the "model." Yet, a joint probability distribution has no inherent directionality. It captures only the likelihood of particular states of variables occurring together, and it is indifferent to which direction any causal influence between these variables runs.[12] If it is entailed by one direction of dependencies between these variables, then it is equally entailed by any dependencies in the opposite direction.

Accordingly, in more recent treatments, the generative model is instead cast as something realized by "the full joint dynamics" of the entire cycle between agent and environment, while the attempt to carve any distinction between generative model and generative process is abandoned (Ramstead et al., 2023). The abandonment of such a distinction seems appropriate in light of active inferences' cyclical nature. It also provides convenient response to Raja et al.'s (2022) objection that free energy minimization simply assumes the existence of a "veridical" generative model. For if this generative model is supposed to be nothing other than the actual dynamics of an agent-environment perception-action loop, then there is no sense in which it could be inaccurate. Still, the removal of any possibility of inaccuracy raises an even more thorny issue: Namely, in what sense do these dynamics constitute a model?

3.4.1 Description versus Instantiation

A possible way to make sense of the FEP's talk of "generative models" as Andrews (2021) suggests, is to interpret this as referring only to a mathematical structure—in this case, joint probability distributions—that need neither be physically realized nor have any representational content. This mathematical structure may subsequently be interpreted as a model of some physical system, but, absent that interpretation, the mathematical structure is not tied to the representation of any particular physical system. Indeed, the same structure may be interpreted to model quite diverse systems. Andrews points to the coupled differential equations of the Lotka Volterra model, for instance, which can equally be used as models of autocatalytic chemical reactions, or of predator-prey dynamics.

That a structure becomes a model only with an interpretation does not mean that this modeling relation is all up to choice, however. Attempting to use the Lotka-Volterra model to predict future demand for vintage beanie babies is likely to lead to poor investment decisions. If an interpretation is to be successful, there must be certain constraints on the structure and dynamics of their interpreted target, in order for these to be mapped to the formal structure of the mathematical equation.

In the case of the joint probability distribution of the generative model, our key requirement is stationarity. A particular joint probability distribution will only be interpretable as a model of a physical process if the behavior of that process is stable, such that its likelihood of being in any particular state will be the same at whatever time we choose to observe it. A pendulum at rest, with occasional perturbations, would be well modeled by such a mathematical structure. The collapse of a building would not.

With this in hand, the natural way to interpret Ramstead et al's (2020) description of the internal arc of the generative process as "instantiating" a generative model is as saying that it has the right sort of dynamics to be described by a generative model. This is to make the positive claims that (1) the relationships between all the parts of our perception-action cycle are stable, and (2) the tendency of any part of this cycle to be in a particular state is stable. While in earlier work this requirement was expressed as the claim that a system is "ergodic" (Friston, 2019b), this has more recently been replaced by the weaker requirement of its being at steady state (Friston et al., 2022; Parr et al., 2022).

If this were not the case, if the system were like a countdown timer, moving in linear progression from one state to the next, then its probability of being in some particular state would be different at time t versus time $t+1$. A fixed probability distribution over the first order states of this system would fail to describe it over time. As mentioned previously, and as will be elaborated in chapter 9, the tool of hierarchical models can allow us to describe what looks like a change in behavior as the preservation of a higher-order statistical regularity, for instance, a reccuring cycle between two semi-stable states, or in terms of the stability of a higher-order state, such as a stable "velocity" with which a timer ticks down.

Either way, as van Es (2021) argues any model here is not something "instantiated" by the process itself. If such a model of the system is instantiated anywhere, it must be in the head of an observer, or in the memory of their laptop. This seems to fit with Ramstead et al.'s (2020) description of the generative model is something that is "defined stipulatively" in order to derive a recognition model that explains subsequent action, and that it "does not have sufficient statistics that are physically realised" (p. 233).

If this were all that is meant by talk of generative models, then our only question would be whether some particular of system that we are interested in has the right sort of stable dynamics to be describable by a joint probability distribution. Yet, this instrumentalist understanding of generative model would fail to support other claims that Ramstead et al. (2020) make such that the generative model is a "control system" that an organism is "equipped with" (p. 231), that it has "causal bite" (p. 233) and is used by the system to "perform policy selection" and to "guide action" (p. 234).

A mathematical structure cannot do any of these things, and systems do not "have," "use," or "perform inferences" with the descriptions that might be given of them. What has causal force is the structure of the system that the generative model describes, not the description itself. A diving gannet may well be describable by a differential equation, but only one of them catches a fish.

As a number of authors have noted, this conflation between the claim that a system is describable by a mathematical structure and the claim that it instantiates a model is one that recurs throughout the free energy literature (Andrews, 2021; Bruineberg et al., 2022; van Es, 2021). Such a presupposition helps to make sense of Friston's unusual way of talking about the existence of models, for instance, where he states that "a model is just an ergodic system" (Friston, 2019b, p. 183).[13] In a more recent paper, Ramstead et al. (2023) explicitly defend this conflation, arguing that "ultimately, there is no way to meaningfully distinguish between saying that the dynamics of a system actually engage in or instantiate approximate Bayesian inference, and saying that they merely 'look as if' they do so." (p. 22).

This conflation is not helpful. At a minimum, a model is a tool used by an interpreter, and inference is an intentional, normatively directed and rationally evaluable process. Models and inferential operations can be better or worse, correct or incorrect. The function of some model may not necessarily be to comprehensively and precisely recapitulate the structure of a target, but a model always has some function. The fact that a system has a particular sort of dynamics merely means that it admits of description by a generative model and mere mathematical describability does not confer any functional role on this property. To say that any stable system is “trying” to infer a probability distribution makes about as much sense as crediting a river-washed pebble with the aim of becoming a sphere or understanding the dynamics of the solar system as an attempt at modeling an orrery. If all there were to being a model was just to admit of a mathematical description, then everything we can talk about is a model, rendering talk of changes as being “caused by a model,” a “process of inference,” or a “model-update” entirely trivial.

Either way, whether the joint probability distribution of the generative model is taken to be literally realized by the dynamics of an agent’s action-perception loop or just to be the statistical description entailed by those dynamics—there is no sense here in which it can be non-veridical. On Ramstead et al.’s presentation, there is no further target, beyond the dynamics that were used to derive this joint probability distribution in the first place, that said distribution is supposed to be a representation of. To introduce this possibility of inaccuracy or inadequacy, we need something else. Specifically, we need an independent basis for postulating a distinct statistical form that an agent’s dynamics *should* conform to, in addition to the statistical description that its actual dynamics entail. Only then would we be able to incorporate the possibility of a discrepancy between said ideal “model” and the probabilistic description entailed by the agent’s actual dynamics.

If we already have a predetermined idea about what kind of behavior an agent should engage in, then describing both these ideal dynamics and its actual behavior in terms of probability distributions allows us to capture the divergence between them. It also induces some constraints on how the former should change to become more like the latter. This is how active inference models of development or learning tend to work, where a system’s trajectory through a series of states that may not recur can be described in terms of a convergence to the steady-state dynamics described by some target generative model. See, for instance, models of morphogenesis as active inference (Friston et al., 2015; Kuchling et al., 2020).

Still, the target generative model here is purely the stipulation of an external observer, and whether some behavior counts as a failure to achieve this target or success at achieving a different target is relative to that observer’s choice. From the low-road perspective, which presents active inference as a description of how an agent achieves goals that have been specified by other means, such an approach is perfectly legitimate. However, if the free energy principle is to ground active inference in an attempt to "explain the intentionality or aboutness of living systems" (Ramstead et al., 2020), then we need it to not only supply a model of an agent's actual dynamics but also a means to derive a generative model of the ideal dynamics that this agent *should* target.

The FEP's attempt to supply a normative principle for generative models will be described in the next chapter. For now, I want to emphasize how this liberal attribution of generative models to systems, based only on their dynamics, impacts how we should understand the claim that a particular system is engaged in active inference.

Firstly, for active inference to get off the ground as an explanation of how a system learns about the causes of its observations, the dynamics of the entire cycle that generates these observations must be stable over time, such that this cycle could either realize, or be accurately described by, a fixed joint probability distribution. That the action-perception loop bringing about an agent's observations has such dynamics is not a given. As I'll argue in chapters 8 and 9, when it comes to a living system, such an assumption holds, at best, only contingently and over a restricted timescale. If this requirement is not satisfied, however, then we do not yet have the means to propose a stable generative model to serve as a target that any internally encoded recognition model might be supposed to approximate.

Secondly, this changes how surprisal should be understood. If surprisal is unlikeliness relative to the generative model and *if* we accept that the generative model just is the actual dynamics of a particular agent-environment system, then surprisal under the FEP is not equivalent to "sub-personal prediction error"—as both (Madary, 2012) and (Clark, 2013) propose in order to distinguish it from personal-level surprise. The surprisal of an event is not something encoded in the brain at all; it is simply the long-run frequency with which this event occurs within a particular generative cycle (Fiorillo, 2010). That something "minimizes surprisal" is just another way of saying that it regularly revisits the same small set of states and that, if it enters into a state that it has rarely visited before, it will quickly leave this state to return to a state that it has more frequently occupied in the past. The minimization of surprisal is not a deliberate process "performed with" a generative model. It is merely a way of describing the dynamics that a process must have in order to be describable by the stable probability distribution of such a model. It is thus not an inferential process in any sense more interesting than the one in which everything is a model.

3.5 Where Does This Take Us, Representationally Speaking?

We've seen that advocates of the free energy principle tend to be pretty liberal in the attribution of generative models, happy to bestow them on any cyclical process with stable dynamics over time. Yet, while Ramstead et al. (2020) are confusingly equivocal on the nature of the generative model(s) of a particular cycle, when it comes to the recognition density of variational inference, they are clear that the internal states of the agent should be taken to literally encode this probability distribution over the "external" variables of the generative cycle.

Insofar as the internal states are a part of the stable generative cycle, then they too must be stable enough be describable by a single stable probability distribution. But why do Ramstead et al. (2020) treat the probability distribution entailed by the dynamics of this subcomponent as a representation of something, while at the same time maintaining that there is no warrant for interpreting the joint probability distribution, entailed by the entire agent-action system, in such representational terms?

Well, insofar as there is a stable joint probability distribution over all the parts of a whole perception action cycle—typically factorized in terms of "external," "sensory," "internal," and "active" variables [P(E,S,I,A)]—so there will be a stable joint probability distribution over any subset of this, such as the joint probability of internal and external states [P(I,E)]. This means that every internal state will correspond to some most likely external state that remains fixed over time. Thus, the idea goes, we can treat each internal states as representing a particular external state (though of course the idea runs just as well in the other direction). We can then rephrase this in the form of probability distributions, such that the probability distribution that describes the behavior of internal states P(I) can be recast as a probability distribution over external states that is "encoded" by the dynamics of these internal states, $P_i(E)$.

As stated, this stable mapping is guaranteed by fact that we can describe our overall system in terms of a generative model, and this holds whenever the system tends to move away from less frequent states to return to the same frequently occupied states—a process we can describe as surprisal minimizing. This means that we can also treat such surprisal minimization as being responsible for preserving this stable mapping between internal and external states. Insofar as surprisal is a component of the free energy equations, we then we can go further and treat that process as a *free energy minimizing* one, and thus redescribe the probability distribution that the internal states entail as a *recognition* model of external states $[Q_i(E)]$, which resembles the actual P(E) as a result of this free energy minimization. Still, insofar as internal states and external states are components of the same coupled cycle with stable dynamics, there should be no difference in the form of $Q_i(E)$ and P(E), no sense in which the former is approximate with respect to the latter. To move from calling this process surprisal-minimizing to describing it as "free energy minimization" may be formally correct, but it is misleading.

Either way, as Kirchhoff and Robertson (2018) argue, all of the above is essentially just a complicated redescription of the correlation between two sets of variables. In a system with stable behavior where all the parts are coupled together in a cyclical pattern of dependencies, such correlation is effectively inevitable. Still, the fact that this correlation is expressed in terms of probability distributions over those variables, and the fact that these distributions become more similar as the divergence between them is reduced via free energy minimization, does not turn this into a representational relationship. All sorts of things may become increasingly similar with time. Over the years, a mattress might start to take on the shape of the person who sleeps there, but the purpose of sleep is not to create a likeness of oneself in springs and foam. Sagging springs are no more a representation of an absent partner than a dry riverbank is of the water that once flowed through it.

Similarity and correlation, like promises of commitment, come cheap. These may be preconditions for a particular structure to be able to play a representational role, but (dis)correlation or (dis)similarity only become (mis)representation if we have independent reason to believe that these structures are being deployed to representational ends. That is, we need to show that the structure is either a constituent of, or used by, a system that is aiming (and thus, potentially, failing) to represent (Millikan, 1984; Dretske, 1993; Ramsey, 2007). With this, we're borne back once again to the issue of function. Only once we've explained the purpose of prediction-error minimization can we answer the question of whether its structures, such as a putative recognition model, play a representational role in service of this.

So, despite Kiefer and Hohwy's (2018)[14] argument that the divergence component of free energy minimization should be interpreted as a measure of misrepresentation, the mere ability to describe parts of a system in terms of free energy minimization is not enough to treat this divergence as a representational, rather than merely correlational, measure. For this, we would need to show that the reduction of this divergence is not merely a fact about what a system does but specifically something that this system is intentionally directed toward achieving, such that we can treat the internal states as having the functional role of representing external states, rather than merely happening to be correlated with them (Millikan, 1984; Ramsey, 2007).

3.6 From Possible Descriptions to Normative Principles

By drawing on the frameworks of variational, and subsequently active, inference, we can see how the process of approximate Bayesian inference can be mapped onto the neural dynamics that PP proposes to identify in the brain. Merely mapping the syntax of (approximate) inference onto a physical system does not prove that the brain *does* engage in inference, however, any more than Chalmers (1995) and Putnam (1975) proved that buckets of water compute addition (Sprevak, 2018).

There are more things zipping and diffusing through our skull than are dreamt of in the ontology of free energy minimization. There may well be plenty of elements connected into the kind of cycles that active inference describes, but there will also be all sorts of random fluctuations and processes of decay or growth that we abstract away from when modeling these stable cycles. To establish a greater legitimacy for the Bayesian in your neural tissue than for the finite state automata in your breakfast cereal, we need to show two things: a reason to privilege the structures picked out by PP and a reason to privilege the mapping between these and the formal structure of active inference.

The natural way to secure this is to show that the dynamical structures singled out by PP are those that explain how the brain performs its function. But, on pain of circularity, we cannot just take the possibility of describing the brain as if it is performing inference to support the claim that inference is the function of the brain and, thus, that this is the *right* description to explain how the brain performs its function. Such an argument becomes even weaker once we see, as we will in chapter 5, the ease with which the free energy framework allows us to trivially map a formal description of approximate inference onto a variety of nonneural systems whose function, if any, it becomes increasingly implausible to regard as inferential.

Still, for all that the free energy framework may allow us to map inferential processes onto a wide variety of systems, it nonetheless induces some quite restrictive conditions on what sorts of systems can be described as free energy minimizers. To even talk about free energy, we need a generative model with a stable assignment of surprisal values for the different states of the system. We can derive this joint probability distribution from the behavior of a system or process only when the tendencies of, and dependencies between, said states have some form of stability over time—that is to say, if our system rarely enters into states that it has not frequently occupied before.

This allows us to hone in on what is meant by claiming that active inference is the function of the predictive brain. Perception may be for action, but now we see that action in turn must be for surprisal minimization and the preservation of a steady state. But now we have to ask, what is surprisal minimization for?

To answer this question, the free energy principle looks to living systems to argue not only for the empirical claim that these do, in fact, tend to instantiate a stable cycle of states but also the normative claim that doing so is a requirement that all living systems *should* satisfy. Under the free energy principle, the preservation of a steady state is not just about epistemological convenience—it's a matter of life and death.

4 One Weird Trick to Stay Alive: The FEP's Philosophy of Life

4.1 The Organism's Agenda

In the previous section, we saw how allowing an agent to actively bring about the very states that it predicts complicates our understanding of the generative process, described by the generative model and approximated by the recognition model. Where surprisal, or free energy, is minimizable through action, so the agent itself becomes part of the "hidden causes" of its own sensory input. As such, the target that our agent attempts to approximate, and minimize surprisal relative to, should not be interpreted as describing the (supposedly) stable dynamics of some agent-independent state of affairs, (as it was on the Helmholtzian characterization) but rather the dynamics of the whole organism-environment system. It is this whole system that makes up the generative process, described by the generative model. Acccordingly, surprisal is neither relative to some internally encoded model, nor just a matter of the objective frequency with which something happens "in general." Surprisal refers specifically to how often an event occurs for some particular organism-environment system. What is surprising for the sperm whale will be depressingly familiar to the bowl of petunias.

So why think that the organism-in-its-environment is a stable surprisal-minimizing process, describable by a generative model and, potentially, approximatable by a recognition one? Well, according to Friston (2012), the process of securing this stability is nothing more than a formal analysis of what it means to be a self-preserving or self-organizing system. "The whole point of the free-energy principle," as he puts it, "is to unify all adaptive autopoietic and self-organizing behavior under one simple imperative; *avoid surprises and you will last longer*" (p. 2117).

So, where variational inference with a recognition model is the ascending shoot of the free energy framework, stretching upward toward a description of inference, action-planning, and other higher-order cogitations, the proposed analysis of self-preservation is where the FEP attempts to root this process in bioenactivist soil, in order to suck biological functions and intentionality up to its cognitive branches. Importantly, Friston does not present active inference as either an alternative to autopoiesis and autonomy or an instrumental means to their end but rather as a first principles analysis of the minimal requirements for existence from which enactivist definitions of life may be derived. So, as Ramstead et al. argue:

> Put simply: active systems are alive if, and only if, there [sic] active inference entails a generative model. This makes the generative model of central importance to the free energy formulation, since it defines the form of life that an organism is seen to enact. (2018, sup. mat. 4, p. 33)

Why think that maintaining a stable generative model by avoiding situations with high surprisal through active inference is the key feature of a living system? A bad way to make this argument involves arguing for the simple imperative as something "circular" or "self-evidently true." (Allen & Friston, 2018, p. 19). In one sense, the claim that all living systems tend to avoid "improbable" states is, indeed, a tautology of existence: things tend to be in states that they are likely to be in. They tend not to be in states that are not likely. It is straightforward to map this on to viability: a state that is nonviable for an organism is very unlikely indeed. It is undeniably correct, though exceptionally pointless, to claim the continued existence of an organism depends on its avoiding states in which it does not exist.

For our laborious excavations beneath the mathematical infrastructure of the free energy framework to result in nothing more than the gnomic utterance that "everything must be what it is, and cannot be what it is not" would be disappointing to say the least. No insight is to be gained by redescribing impossible states as improbable ones, then sagely noting that an organism seeking to continue its own existence must avoid these.

Besides, as a way of staying alive, minimizing the surprise of your own death would come one action too late. Taking surprisal to be a binary matter of "possible, or not" would indeed make the claim that it *is* avoided tautologous, but ought implies can. To say that minimizing surprise is something that organisms *must* actively do implies that they *can* be in some relatively high-surprisal states such that they can then move away from them back to more likely ones.

As it is presented in the FEP, surprisal is crucially something that comes by degrees, such that it can be minimized. Specifically, as discussed in section 3.4, how much surprisal some state has is determined simply by how frequently it has occurred in the history of our target system and to say that an organism minimizes surprisal is to say that it frequently revisits the same small set of states. As Friston (2018) puts it,

> We are only interested in one sort of system. These are processes where (the neighborhood of) certain states are re-visited time and time again; for example, the biological rhythms that characterize cardiorespiratory cycles—or the daily routine we enjoy every Monday, on getting up and going to work . . . This means, on average, I must move toward states I am more likely to occupy. This may sound trivially simple but has enormous implications for the nature of any (interesting) process that possesses an attracting set of states. (p. 2)

So, rather than a tautology derived from first principles about what it means for something to exist, the FEP looks more like a positive proposal about the particular kind of existence living systems have. The empty platitude that an organism is more likely to be in states that are more probable becomes the substantive argument that (1) the organism's states vary and (2) it tends to repeatedly revisit the same set of states it has visited previously and to avoid those states that it has not previously visited with any great frequency.

While the state of our system may constantly fluctuate, the probability distribution over these states must remain invariant such that we can treat this invariant probability distribution as the generative model of that organism. This is all there is to saying that an organism minimizes surprisal and thus that it "entails" a generative model. To say that an organism's self-preserving behavior necessarily "entails" a generative model, in this sense, means only that it will be describable by a joint probability distribution that does not change over time. It does not establish that any part of this organism, brain or otherwise, literally encodes a separate recognition model. The latter claim is a suggestion about how the brain *could* perform inference about the generative process, given that the exact generative model (as opposed to the approximate recognition model) would be computationally intractable. But the need to attribute such a model depends on the assumption that organisms need to explicitly encode and compute with some model of the generate process in order to guide their actions. As discussed in the previous chapter, statistical properties of an organism's dynamics alone are not enough to substantiate this.

As such, in papers that focus on the free energy framework as a theory of biological self-organization, rather than on the specific problem of how the brain might actually perform approximate Bayesian inference, there is no role for a separate recognition model (Friston, 2013). Somewhat misleadingly, however, Friston and colleagues continue to speak in terms of "free energy" and to describe the organism's behavior as "free energy minimization." This is technically correct in as much as free energy is *divergence* (the KL-divergence between the recognition and generative model) plus *surprisal*, and so when there is no separate recognition model and no KL-divergence, it reduces to surprisal alone. Still, in as much as "free energy" implies the use of variational inference and the existence of an approximate model, encoded by the organism and distinct from the generative model that describes organism-environment dynamics, so the continued use of the term where no such encoding has been established is misleading (Mann et al., 2022). While I would prefer to use "free energy" only in contexts where we have these two distinct types of models, the use of it to mean nothing more than surprisal relative to a generative model is so pervasive as to be unavoidable. It will have to suffice to emphasize that, going forward, any mentions of "free energy" mean nothing more than "surprisal" and "surprisal," in turn, means nothing more than the infrequency of an event within a particular system.

4.2 Self-Organization and Steady State

In a series of papers, Friston (2013, 2018, 2019b) argues that the stability-based account of what it is to be a living system follows directly from a definition of an agent in terms of the properties of ergodicity, and low entropy. This is most clearly expressed in Friston and Ao (2012), where they state that

> for a well-defined agent to exist its ensemble density must be ergodic; that is, an invariant probability measure. In other words, the density cannot change over time; otherwise, the definition of an agent (in terms of the states it occupies) would change. A simple example here would be the temperature of an organism, whose ensemble density is confined to certain phase-boundaries. Transgressing these boundaries would change the agent into something else (usually a dead agent). (p. 2)

So, it is this requirement of ergodicity that is initially used to mandate that the generative model, that is, the distribution over possible states, will be invariant over time. The "low-entropy" part is not used in the thermodynamic sense but rather in the information-theoretic sense of Shannon (1948). It specifies that the particular states our system revisits will be only a small subset of all possible states, such that the probability distribution over possible states has low variance/high precision.[1]

We need the requirement of low entropy, for ergodicity is not enough to entail surprisal minimization. Repeatedly rolling a six-sided die is an ergodic process, but if it is a fair die then each side is equally likely, its entropy is at maximum, and its stable generative model would just be a flat distribution over all possible states. For such a maximal entropy system, any sequence of states would have the same surprisal, and so surprisal is not minimized.

Unfortunately, ergodicity actually implies more than stationarity of dynamics over time, a point that has caused some controversy in recent work on the FEP (Palacios & Colombo, 2021). Strictly speaking, ergodicity requires not only that the average behavior of a particular iteration of a system will be invariant over time but also that it will be insensitive to initial conditions. Take a spinning top untroubled by external perturbations. Left alone after an initial impulse, this would quickly settle into orbiting a small area of the overall tabletop—a stable regime, describable by a stationary probability distribution over its position. Which particular part of the table it ends up oscillating around will, however, differ depending on the initial impulse and starting point. As such, spinning a top is not an ergodic process.

This stronger requirement of ergodicity has a few important consequences. Firstly, it implies that any particular iteration of a system will eventually visit every state that it is possible for that system to inhabit. Secondly, it implies that a snapshot of an ensemble of iterations at any single point in time will converge with the distribution across states for the trajectory of a single iteration *over* time, as the duration of the individual trajectory or the size of the ensemble increases. This property is expressed in Birkhoff's (1931) ergodic theorem, that with increasing samples and increasing time, the ensemble average and time average will eventually converge.

The classical example of ergodicity in statistical physics is that of idealized gas particles bouncing around a container. Idealized is the key word here for, as Palacios and Colombo (2021) note, proving the existence of concrete systems that meet this requirement has been extremely difficult. In many physical systems, the time required to exhibit every possible configuration, or for the time average to converge, extends far beyond the duration over which the system exists (Palacios 2018; Gallavotti, 1999). It is now, they claim, widely recognized that most of the systems studied in statistical mechanics are most likely non-ergodic (Earman and Rédei 1996; van Lith, 2001).

This might be to the advantage of the FEP if it were shown that such ergodicity is the distinctive preserve of the biological. Unfortunately, the opposite appears to be the case. Ergodicity is even more implausible in the biological sphere. To take the favored example of Stuart Kauffman (2000), who argues that it is precisely the defiance of ergodicity that defines biological organization, it would take 10^{39} times the current life span of the universe to make all possible permutations of a 200-amino-acid-long protein at least once. The convergence of an ergodic process cannot be responsible for the stability of the specific subset of amino acid combinations we observe. As Kauffman puts it:

> It follows that, even if we consider the universe as a whole, at the levels of molecular and organizational complexity of proteins and up, the universe is kinetically trapped. It has gotten where it has gotten from wherever it started, by whatever process of flow into a persistently expanding adjacent possible, but cannot have gotten everywhere. The ergodic hypothesis fails us here on any relevant timescale. (2000, p. 145)

An ergodic system forgets its history. No matter where it starts, after enough time, any iteration of an ergodic process becomes indistinguishable from every other. Biological processes, however, fall into local stability wells–where they start matters for where they end up. Any plausibility that the assumption of ergodicity might have for theoretical gas particles bouncing around a box is utterly lacking for the specificity and variation of living organisms.

The ways in which biological processes defy ergodicity are fascinating and informative as to the distinctive character of living systems, and I will discuss them in chapters 8 and 9. For now, I suggest we simply accept the move made in Friston (2019a) and Da Costa et al. (2021) to shed the unnecessarily strong claim of ergodicity and limit ourselves to merely requiring that the probability distribution for each particular iteration of a system remains stationary over time—in dynamical systems terms, the requirement that it reaches a *steady state*.

Importantly, this requirement does not amount to the claim that the organism only stays in one state. It may fluctuate regularly from a single state or cycle around several states. Moreover, the steady "state" here can be understood quite liberally to include "states of motion" such as velocity or acceleration—in which case what remains steady might be a "rate of change" or some higher-order derivative of the same. I will explore the possible ways that this steady-state requirement might be understood in chapter 8, to argue that many of the objections raised against the FEP's assumption of ergodicity will also apply equally to this weaker requirement.[2] For now, in the interest of getting the whole theory onto the table before we begin dissecting it, I will temporarily put these aside.

So, let's just say I am indeed defined by being a system at steady state and you want to know what I'll be up to at any point in future. You could observe my behavior over the course of several weeks and chart the relative frequency with which various states are visited. You'd find a very small subset of states—making coffee, drinking coffee, staring mournfully at a mug that is now empty of coffee—to be repeatedly revisited with a high frequency.[3] The vast majority of other, in principle entirely possible, states—relaxing with a glass of champagne in the Balmoral Bar, executing a flawless underwater handstand at the bottom of St. Margaret's Loch, spinning fire on the Meadows with the Beltane Society—will be occupied extremely rarely if at all.

Or as Friston et al. (2020) put it:

> At a larger timescale, this trajectory could reflect your daily routine, getting up in the morning, having a cup of coffee, going to work and so on . . . The key aspect of this trajectory is that it will—after itinerant wandering and a sufficient period of time—revisit particular regimes of state space. (Friston et al., 2020, p. 31)

If I am defined by being a steady-state system, then once you have identified this limited set of states, you have my behavior sussed for life. Why think that this is true? The idea that my behavioral tendencies will not change or evolve over time seems as unlikely as it would be depressing. The more cynically minded might suggest that when all you've got is a joint probability distribution, then everything looks like a steady-state system. But the FEP is not the first to advance stability as a principle of survival. Before I get to the critical half of this book, in chapter 7 onward, it's worth looking at what motivates the link between biological survival and stability of dynamics.

4.3 Cybernetics Redux

Temporarily postponing immediate philosophical and emotional objections to the idea that such monotony delimits my entire behavioral repertoire—there is at least something to this stability as a common tendency of life as we know it. As both Colombo and Wright (2021) and Seth (2015) argue, this "simple imperative" of avoiding surprises is foreshadowed in the work of the early cyberneticist W. R. Ashby, who sought to provide an analysis of the adaptive, self-organizing behavior of living systems that meets the following criteria: "(1) it is purely objective, (2) it avoids all metaphysical complications of 'purpose,' (3) it is precise in its definition, and (4) it lends itself immediately to quantitative studies." (Ashby, 1940, p. 479)

Ashby's solution for this was the theory of "generalized homeostasis." By reducing survival to a matter of stability, he proposed that the mysterious appearance of goal-directed behavior in living systems—no matter how complex and unexpected—results from nothing more than the same tendency to return to equilibrium when perturbed that is exhibited by *all* stable physical systems. As he proposes in *An Introduction to Cybernetics:*

> Thus the concepts of "survival" and "stability" can be brought into an exact relationship; and facts and theorems about either can be used with the other, provided the exactness is sustained. The states M are often defined in terms of variables. The states M1, . . . , Mk, that correspond to the living organism are then those states in which certain essential variables are kept within assigned ("physiological") limits. (Ashby, 1956, p. 197)

And thus, as he claims in a later paper:

> We have heard ad nauseam the dictum that a machine cannot select; the truth is just the opposite: every machine, as it goes to equilibrium, performs the corresponding act of selection. Now, equilibrium in simple systems is usually trivial and uninteresting; it is the pendulum hanging vertically; it is the watch with its mainspring run down; the cube resting flat on one face . . . What makes the change, from trivial to interesting, is simply the scale of the events. (1962, p. 270)

Situated as a revival of the (less mathematically abstruse) project of Ashbyian cybernetics, the meaning of the simple imperative of the free energy framework becomes clearer. Talk of minimizing surprisal translates to countering the deviation of an essential variable from assigned limits, and to say an organism must avoid surprises just means that it must maintain homeostasis of its essential variables.

Like Friston, Ashby's work goes beyond the submission of adaptive behavior to formal analysis. Similarly, he sought to account for our cognitive operations in terms of how they serve the coordination of such behavior. Defining survival as stability not only cleared the ground of teleology or purpose but also laid the foundations for the analysis of the brain as a control system tasked with maintaining the stability of our essential variables. This should not be interpreted as the transparently false claim that our entire neural architecture is dedicated solely to the triggering of autonomic reflexes. Rather, as Ashby declared, his intention was to show "how all the organism's exteriorly-directed activities—its "higher" activities—are all similarly regulatory, i.e., homeostatic" even where what is regulated need not necessarily be a bodily state (Ashby 1956: 196). The manner in which he proposed to do this was both ingenious and somewhat perplexing.

What distinguishes the organism, he argued, is an additional mechanism, found only in a subclass of "ultrastable systems." The role of this mechanism is to trigger the random reorganization of the structure of the system when its essential variables are pushed beyond the threshold that it has either "adapted," or been designed, to compensate for. In the organism, this reorganization amounts to changing the parameters of the behavior-producing network at random—a process that ceases only when a behavioral policy is discovered that brings its essential variables back to stable equilibrium.

When Ashby presented a working model of this, called the homeostat, his "electronic brain" generated both headlines in the popular press (Ashby, 1949) and bemusement from his fellow cyberneticians. Julian Bigelow summed up the general attitude in declaring that "it may be a beautiful replica of something, but heaven only knows what" (Husbands & Holland, 2012, p. 12). Barring the supposition that Ashby was a particularly unorthodox chess strategist, it was far from obvious how a system that flails around randomly until its stability is restored might one day, as he proposed, play the game "with a subtlety and depth of strategy beyond that of the man who designed it." (Ashby, 1948) Such a task seemed more immediately achievable by the symbol processing systems of Ashby's contemporaries, Herbert Simon and Allen Newell, which thereby came to define the dominant paradigm in AI research for many years after.

Still, as Vernon (2013) emphasizes, while the homeostat and the concept of ultrastability may have represented the apotheosis of Ashby's working models, his theoretical speculations in "Design for a Brain" extended further to the discussion of metastability: as a property that emerges from a high number of interconnected ultrastable systems, and one that might better characterize the operations of something as complex as a nervous system. A specific model of how the nervous system might function in this regard was later developed in *Behaviour: The Control of Perception* (1973) by the cybernetician William T. Powers, wherein he describes how hierarchical regulation of sensory variables could indirectly support the regulation of external variables in order to thereby maintain the stability of essential variables.[4]

This sounds extremely similar to the operation performed by a predictive processing brain. Unfortunately, the preference for the language of "predictions" and "errors," over "reference levels" and "discrepancies" in the FEP (and consequently in predictive processing), has obscured these similarities by aligning the latter accounts with the Bayesian brain tradition, over its cybernetic ancestor. Once the interchangeability of these terms under the FEP is made clear, the similarities are striking. If the simple imperative of surprisal minimization is simply a revival of Asbhy's reduction of survival to stability, then we can frame active inference—and the predictive processing implementation of this—as providing an analogous proposal to Powers's perceptual control theory. An explanation of how, by taking sensory variables as proxies for essential ones, and by attempting to predict and control these over multiple timescales, a system can engage in the model-based regulation of its environmental interactions, anticipating and countering dangerous tendencies before they threaten its core stability.

There is, regrettably, not the space to engage in an extended comparison of PP and perceptual control theory here. For now, it should suffice to note that despite their many similarities, there remains (at least) one novel element to the FEP. If the theory is in tune with both Bayesianism and cybernetics, then it offers a potential bridge between the latter's language of control and the inferentialist vocabulary of the former. The value of the FEP might then be in its ability to integrate both the bio-physical grounding of cybernetics in feedback control, and the abstract forms of rationality modeled by Bayesian Brain frameworks, under one unifying formalism.

4.4 Stability and Agency

If comparing Friston's project with that of Ashby's helps bring clarity to the free energy framework, and it also brings into focus tensions within it. While Ashby is often credited with popularizing the term "self-organization," he is nonetheless at pains to emphasize that there is "no such thing"—at least not when this is understood with volitional implications, as the result of some drive that originates from within the organism itself (Ashby, 1962). All such behavior, he argues, can only be explained if we view it as a response to external causes. When environmental perturbation drives the system away from a stable state, it triggers this system's inevitable return to equilibrium. The stable system, the ultrastable system, and the multistable system are all driven by the same laws and forces; the only difference is in the degree of mechanical complexity.

For Ashby then, the whole point behind the concept of *generalized* homeostasis is to strip biological explanations of teleological language, by proposing that the release of glucose stores to counter a decrease in blood-sugar levels, the onset of shivering in response to a drop in body temperature, a pendulum falling back into kinetic equilibrium, and the determined uprightness of the wobble-doll differ only in degree of robustness to perturbation, not in kind. The widespread failure to identify this continuity, he suggests, can be attributed to the lack of systems of intermediate complexity between the multistable person and the basic pendulum. "The computer," he argues,

> is heaven-sent in this context, for it enables us to bridge the enormous conceptual gap from the simple and understandable, to the complex and interesting. Thus we can gain a considerable insight into the so-called spontaneous generation of life by just seeing how a somewhat simpler version will appear in a computer. (Ashby, 1962, p. 271).

For Ashby, the stability of key variables does not just define survival and the persistence of organism but a general tendency of all physical things. Similarly, Friston (2019a, 2019b) sometimes presents the free energy principle as having such generality, describing the requirement that the states of a system have steady-state dynamics as part of an all-encompassing definition of existence by means of which "everything of interest about life and the universe can be derived" (2019b, p. 176).

The idea that stability is an important principle for more than just biological homeostasis runs as follows: if the region of possible states of some system were constantly changing, then we would have no way to reidentify it over time. If the region were stationary but did not have low entropy, then a wide range of states would be equiprobable at any point and there would be no distinctively characteristic states by which this particular system could be identified. As such, Friston (2019b) presents these criteria as amounting to necessary constraints on the possibility of ongoing existence.

The thought that this might serve as an exhaustive definition of all forms of existence seems, to me at least, wildly unconvincing. Nonetheless, once we have this putative definition, we can see why Friston often describes the free energy principle as tautological, for when existence is defined in such terms, anything that exists will end up satisfying it.

What makes Friston's and co-authors' claims about the free energy principle so confusing is that, despite sometimes appropriating this Ashbyian analysis of the generalizability of homeostasis, at other points, they have instead presented homeostasis-as-stability to be something *distinctively* biological, evidencing a previous lack of appreciation for its generalizability to all sorts of inanimate systems. As Friston (2010) states,

> The defining characteristic of biological systems is that they maintain their states and form in the face of a constantly changing environment [. . .] This maintenance of order is seen at many levels and distinguishes biological from other self-organizing systems; indeed, the physiology of biological systems can be reduced almost entirely to their homeostasis. (p. 127)

And, as Badcock et al. (2019) put it,

> This is the remarkable fact about living systems. All other self-organising systems, from snowflakes to solar systems, follow an inevitable and irreversible path to disorder. Conversely, biological systems are characterised by a random dynamical attractor—a set of attracting states that are frequently revisited. (p. 3)

Where Ashby cautions against the use of volitional language to characterize "self-organizing" systems, discussions of the FEP, in contrast, make liberal use of agential and teleological terminology in describing surprisal minimization. This habit persists even in light of more recent extensions of the free energy principle beyond the biotic sphere, to the realm of physical mechanisms in general. Thus, we find Ramstead et al. (2023) declaring that:

> We can view the NESS [steady state] density as providing a set of *prior preferences* which the particular system looks as if it attempts to enact or bring about through action. Indeed, we can think of this solution to the dynamics as a naturalized account of the teleology of cognitive systems. (p. 10)

When deployed exclusively in relation to neural dynamics, the claim that surprisal minimization describes an "imperative" that an ergodic or steady-state "agent" must follow to "actively maintain" itself (Friston, 2013) does not immediately strike one as obviously misplaced. Yet, the free energy framework is supposed to apply to all biological systems, not just en-brained ones. If it does so only in as much as surprisal minimization describes the stability-through-fluctuation of *any* old physical system, then the idea that it licenses talk of "active maintenance," following "imperatives," or even "inferring" and "modeling" becomes much less convincing.

So, surprisal minimization is too general to distinguish the unique characteristic of animate existence. Neurocentric predictive processing, conversely, is too narrow. This is a major issue for the FEP as a principle of biological self-organization, and an even worse blow to pretensions at supplying a bioenactivist theory that takes the distinction between life and non-life as fundamental.

Despite its significance, this concern over whether the FEP has the required specificity to serve as a theory of living agents has only recently begun to meet with serious critical engagement in the free energy literature (DiPaolo et al., 2022; Raja et al., 2021). As such, I think it's fair to characterize the responses to this challenge as at a relatively nascent stage of development. Still, at the time of writing, two different strategies can be distinguished. The first is to accept that a realist account of the intentionality and autonomy that differentiates living systems cannot be derived from the requirements of steady-state dynamics or surprisal minimization. The second involves attempting to derive a principled distinction between different kinds of steady-state surprisal minimizers that corresponds to the distinction between life and non-life.

The first response to these issues is to simply accept the generality of the FEP, that is, to accept that mere steady-state dynamics cannot suffice for attributing intentionality to a system. Many who have elsewhere defended the view of the FEP as an explanation of biological autonomy have now moved in this direction, and FEP-derived attributions of "belief" "inference" "action" are increasingly prefixed with the caveat "as if," where the FEP is understood as merely describing the dynamics a system should have to look *as if* it is intentionally directed (Ramstead et al., 2023; Friston et al., 2020).

Unfortunately, however, this change of interpretation is not always explicitly acknowledged or consistently maintained. In the same paper, Ramstead et al. (2023) assert the vacuousness of the FEP as something that does not allow us to "construct a useful demarcation between bona fide cognition and dynamics appearing merely 'as if' they are cognitive," (p. 21) before immediately going on to claim that, nonetheless, everything that can be modeled as if it performs inference can be understood as actually performing a sort of inference. This rejection of any distinction at between genuine and "as if" attributions of inference could be understood as an eliminativist position were it not for their claim that such a move provides us with a "genuine teleology of self-organizing systems" (p. 21).

We cannot both have our teleological cake and eliminate it. To both preserve a genuinely inferential teleology and to deny that it is distinct from "as if" inference does not deliver instrumentalism or eliminativism, but rather pan-inferentialism—a position far from their proposed "restorative" abandonment of "metaphysical baggage of statements about 'mind' and 'cognition'" (p. 21).

Similarly, despite proposing an instrumentalist interpretation of free energy minimization in Friston et al. (2020), Wiese and Friston elsewhere continue to maintain that living systems do genuinely have the teleological function of minimizing free energy and that this accounts for the emergence of autopoiesis and the emergence of goal-directedness—with no "as-ifs" to be found. As they state,

> In other words, the FEP provides a (very general) answer to the question "What *does* a system do, when it stays alive?" as well as to the question "What *should* a system do, in order to stay alive?" The FEP thereby accounts for a basic form of goal-directedness that is characteristic for basic forms of intentionality. (Wiese & Friston, 2021, p. 7)

Still, putting aside this lack of consistency on whether we are dealing with bona fide intentionality, or just the appearance of it, I think the transition toward instrumentalism about the FEP's constructs is exactly the right move to make. The presentation of a general property of all sorts of systems, which can serve as a useful constraint for model-building, may be less seductive than a first principle for explaining the emergence of life and intentionality. Still, accepting this more modest framing positions us for the more productive program of attempting to explain appearance of a difference between life and non-life, between agency and inanimacy, using the tools of the free energy minimizing framework.

One strategy is to appeal to the increasing levels of hierarchical complexity that can be modeled under active inference, which might allow us to identify a subclass of systems that can minimize free energy with respect to future events and probabilistically weighted counterfactuals. Such an approach suggests that it is this increase in the complexity of surprisal-minimizing mechanisms that differentiates between simple homeostatic systems that should merely be modeled "as if" they are engaging in active inference versus allostatic mechanisms that might qualify for the application of psychological and intentional predicates like "belief," "action," or "perception" (Friston et al., 2020).[5]

In taking the question of whether or not a system is an intentional system, a living agent, or a genuine active inferrer to be a matter of degree, such a response is thoroughly Ashbyian. In reducing the distinction between life and non-life to a practical choice about where is most convenient to draw a line on this continuous scale, this gradualism is fundamentally incompatible with our bioenactivist goal to provide a naturalistic, and thoroughly *realist*, account of basic intentionality.

Still the fact that the FEP cannot deliver a bioenactive account of autonomy does not necessarily mean it is a bad account of living systems. Perhaps Ashby was right. Perhaps the enactivist is wrong to see differences in kind where there are only differences in degree. To assess the prospects of this Ashbyian version of the FEP, however, we will need a means to analyze nonneural lifeforms, such as single cells, as active inferrers. This requires more than just steady-state dynamics, but also a means to factor the system into active and sensory components. To see how the free energy framework attempts to extend such an analysis beyond the brain, we will need the additional tool of a Markov blanket, which can be deployed to demarcate such a sensorimotor interface. This will take some explaining, and so I'll put this strategy aside until chapter 5.

That the applicability of the FEP does not mark a difference in kind between life and non-life does not mean the only option is to look for differences of degree *between* free energy minimizers. Another (at this stage somewhat underdeveloped) proposal suggests that there is in fact a dichotomous division within the class of free energy minimizers, one that is not merely a matter of where we choose to draw a line. This strategy starts with the common observation, most notably attributed to Schrödinger (1951), von Bertalanffy (1968), and Prigogine and Stengers (1984, 1997), that organisms persist out of equilibrium with their environment and attempts to characterize this in terms of a specific form of steady-state dynamics. As this does not rely on the apparatus of a Markov blanket, and, as I believe it is not successful, I will briefly discuss it here.

4.5 The More Things Change, the More They Stay the Same

The FEP, as we've seen so far, has been proposed as an account of steady-state systems, in the specific sense of systems where the probability distribution over possible states remains stationary over the duration of the system's existence. Unlike the stronger requirement of ergodicity, this allows that the system can be sensitive to initial conditions such that the particular steady state that it eventually settles into may vary across different iterations.

So far, I have discussed this in the context of a homeostatic set point where the steady state is a fixed-point attractor, such as a body temperature of 36.5 degrees, from which the system only departs as a result of random fluctuations and to which it reliably returns. This kind of point attractor is what we would see when the system's dynamics are exclusively driven by what is termed the "dissipative" (or curl-free) flow back to more likely states, which counters surprisal-raising fluctuations in order to prevent the dissipation of the generative model. If random symmetrical fluctuations are all that perturb the system, then this flow alone is sufficient for the maintenance of a steady state and is the only thing actually *mandated* by the FEP's "simple imperative" for living systems.

Yet, even in the case of simple homeostatic processes, like thermoregulation, what we actually find is not a set point but a circadian rhythm, a recurring cycle through different, equally viable, temperatures. Fortunately for the free energy framework, while the return to a single fixed point is sufficient to secure the preservation of a stable generative model that defines a steady state, it is not the only way to do so. As Friston and Ao (2012) discuss, the preservation of a stationary probability distribution is also *compatible* with (though does not require) the system's dynamics also having a second component, called the solenoidal (or divergence-free) flow. This can be extracted for any system at steady state via the Helmholtz decomposition, which breaks the system's overall flow into both dissipative and solenoidal components.

Rather than driving the system back to more likely states, as the dissipative flow does, the solenoidal part circulates around a number of equivalently likely alternatives. This component of the system's dynamics does not itself reduce surprisal; it serves neither to dissipate nor to counter dissipations of the generative model. It is merely compatible with a generative model remaining unchanged. The dynamical pattern of systems with a solenoidal component will thus be a cycle around a limited subset of states, for example, the orbit of a planet, or a circadian rhythm, rather than a single point attractor.

Strictly speaking, the FEP requires only the dissipative flow needed to minimize surprisal-raising fluctuations and to preserve a stable generative model. While a solenoidal component is compatible with this, it is not required for a system to be free energy minimizing. Nonetheless, the possibility of incorporating this solenoidal component has been suggested to distinguish systems at non-equilibrium steady state (Friston, 2019a; Friston et al., 2021). The idea is that if we can identify a part of the dynamics of a system's internal states that is not driven by response to perturbations (as the dissipative-flow is), then we can identify the "intrinsic dynamics" by which this system, as an autonomous one, keeps resisting the descent into equilibrium balance with its environment.

While this solenoidal flow is not much discussed in Friston's (2013) paper, in which he sets out the FEP as an account of the emergence of life, in a revealing aside, he suggests that it may be key to distinguishing the out-of-equilibrium nature of organic systems. As he hints:

> However, minimum entropy is clearly not the whole story, in the sense that biological systems act on their environment—unlike a petrified stone with low entropy. In the language of random attractors, the (internal and Markov blanket) states of a system have an attracting set that is space filling but has a small measure or entropy [. . .]. Put simply, biological systems move around in their state space but revisit a limited number of states. This space filling aspect of attracting sets may rest on the divergence-free or solenoidal flow (equation (2.3) that we have largely ignored in this paper but may hold the key for characterizing life forms. (Friston, 2013, p. 11)

This suggestion continues in Friston (2019) and in Ramstead et al. (2023), where it is argued that the solenoidal flow underpins the kind of exploratory behavior associated with living systems. Still, given that all sorts of non-biological systems, from the hydrological cycle to planetary orbits, have a solenoidal component to their dynamics, this component does not, as the above authors imply, seem to take us that far in getting at what is distinctive about living systems in particular. Further, as Aguilera et al. (2022) argue, incorporating this solenoidal flow appears to require prohibiting the kind of patterns of influence needed to formalize a "sensorimotor" interface in terms of conditional independencies between parts of our organism-environment system—which will be needed when we attempt to extend free energy minimization and active inference beyond the brain in chapter 5.

Either way, the details of the ability of the FEP to formalize non-equilibrium dynamics does not affect the fact that it has been presented as a principle of systems that can be characterized by a stationary probability distribution. As a theory of living systems, the FEP requires that the probability of their being in any particular state stays constant over time. This underpins claims of a supposed "surprisal-minimizing" imperative for the organism. As such, I will put aside the specific issue of non-equilibrium status, which appears as a somewhat ad hoc supplement, to stick with the weaker requirement of a system that converges to any sort of steady state. That is, any system that can be described by a stationary joint probability distribution—or, in FEP terms, that "entails a generative model."[6]

In the next chapter, I will consider efforts to extend this beyond the nervous system, in the attempt to characterize a process that is not brain-bound but is still intended to be distinctively biological. After showing how such a characterization still fails to achieve the required level of specificity for a theory of life, I'll discuss how work on the free energy principle has responded in a more recent turn toward an Ashbyian view on the continuity between the organic and inorganic. Such a turn, I will argue, not only invalidates the FEP's enactivist credentials but also reveals its inability to deliver the first principles of living systems.

5 Free Energy Minimization beyond the Brain

Minimizing surprisal just means maintaining stability, and maintaining stability means surviving, but how exactly does this noble goal hook back up to the messy business of active inference? Active inference may be a means of minimizing surprisal, but this description of a system does not just drop fully formalized from the observation that the system is surprisal minimizing (or stable, homeostatic, or ergodic) alone. In addition, we need our system to be broken down into specific sets of variables, namely: internal states, external states, sensory states, and active states, such that we can map these to the specific generative cycle proposed under active inference.

Thus far, this decomposition has been presented to us ready-made in the structure of the nervous system and its sensorimotor interface. If the goal of the FEP was no more than the specification of a procedure by means of which brain-enabled creatures like us can maintain our viable states, then we could stop here, consider ourselves satisfied with this as no small accomplishment, and go out and play in the afternoon sun.

Regrettably, FEP's advocates are of a more ambitious bent. Since its (comparatively) modest origins as a "theory of cortical responses" in Friston (2005), these imperial aspirations have driven the FEP's expansion in scope, sights now set on the provision of "a mathematical formulation of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder" (Friston, 2010, p. 127), and one that can be applied to all forms of life, "from single-celled organisms to social networks" (Friston 2009, p. 293).

This raises the question: Where exactly do we find the sensory, active, and internal states of the *Myxogastria* slime mold, the flash mob, or the Bank of England's Monetary Policy Committee?

To understand the FEP's answer to this, we need to get one further piece of mathematical machinery on the table. This is the Markov blanket, introduced by Judea Pearl (1988) and, in its original form, it is nothing more than a representation of statistical separation between nodes in Bayesian networks. In the FEP, this was initially introduced to separate out those variables whose state becomes informationally irrelevant with respect to the state of some target variable once we know the state of this smaller subset of nodes. This factorization, as we saw, is exactly the kind of thing that needs to happen in order to implement approximate Bayesian inference via hierarchical predictive processing. Indeed, Markov blankets first appeared in the free energy literature in precisely such a guise (Friston et al., 2007).

In Friston (2013), however, Markov blankets are deployed as a means to identify real-world boundaries, such as the separation between the internal states of an organism and the external state of the environment, and further to partition this boundary into the active and sensory states of the active inference equations. As a result, this unassuming construct has been pressed into a series of increasingly demanding labors, from establishing an *epistemic* boundary (Hohwy, 2016, 2017) to formalizing an *autopoietic* boundary that is both produced by and preserves the system that it so bounds (Palacios et al., 2017: Friston, 2013; Kirchhoff et al., 2018; Allen & Friston, 2018). This has induced such a degree of conceptual hypertrophy that Bruineberg et al. (2022) suggest dubbing this new construct a "Friston Blanket" to distinguish it from its humble ancestor.

The claim that a Markov or "Friston" blanket allows us to map active inference equations onto a diversity of systems beyond the brain, together with the claim that said blanket formalizes the autopoietic boundary of biodynamic enactivism are thus both crucial steps in the idea that the FEP might provide this "first principle" of living systems. If, in contrast, the individual biological agent must already be demarcated prior to our being able to approach it in terms of free energy minimization, then the FEP cannot be our *first* principle for biological autonomy.

To evaluate whether Markov blankets are capable of playing either of these roles then, we need to go back to their origin in Pearl's work on Bayesian networks. This will, unfortunately, require a rather prolonged detour from our aim of understanding the FEP as an analysis of living systems, due to the extent to which the notion of a Markov blanket has been distorted in the FEP literature and the lack of explicit acknowledgement of these distortions. If you are already familiar with Markov blankets within Bayesian networks, however, you can skip this section and proceed straight to section 5.2.

To claim that the Markov blanket is any sort of boundary for the organism itself—epistemic, autopoietic, sensorimotor, or otherwise—the free energy framework has presented these blankets as part of the real world and not just a feature of our models of this world as they were initially proposed. This claim has typically been interpreted as a regrettable mistake of, as Andrews (2021) puts it, "confusing the math for the territory" (see also Bruineberg et al., 2022; Menary & Gillet, 2021; Beni, 2021; and Raja et al., 2021). Nonetheless, while Friston and colleagues may not provide any justification for the metaphysical claims of Markov blanket realism, I don't think that it is *necessarily* a mistake. There is, as I elaborate in the appendix, a positive position one could take on the structure of reality and the metaphysics of causation such that it would be plausible to take Markov blankets as real entities, not just modeling constructions.

Now I'm not saying that such a position is very compelling. Either way, it doesn't really matter because, as I will argue in the next chapter, the main problem concerning Markov blankets is less whether they are "real" or "constructed" and more that they are neither an essential feature of organisms nor of any description that can capture the necessary conditions of their existence.

5.1 A Brief Review of Causal Inference and Markov Blankets

A Bayesian network is a means of representing probabilistic relationships via a directed acyclical graph (DAG for short) in which variables (each represented as a node) are connected by directed lines between, representing their direct statistical dependencies (Pearl, 2000, 1988; Spirtes et al., 2000) (see box 5.1). "Graph" here is just the mathematician's term for "network." A DAG is directed in that the connections between nodes have a particular direction of influence, and acyclic in the sense that this does not circle back on itself.[1]

To create a Bayesian network, all that is needed is a joint probability distribution (such as the generative model of the FEP) over the values of our variables and a few simple axioms of construction. The most straightforward reason one might do this is to visualize the decomposition of an unwieldy joint probability distribution over a large number of variables into a set of more tractable independent distributions over smaller subsets of variables. We can then select one of these depending on which particular variables we are concerned with. To illustrate how a Bayes net is constructed, we can take a simple set of three variables X, Y, and Z.

If $\mathrm{P}(x|y) \neq \mathrm{P}(x)$, that is, if knowing that $Y = y$ changes the probability that $X = x$, then we have a statistical relationship between the two. This is not enough to justify drawing a line directly from Y to X yet, however. To see why, suppose it is also the case that, knowing that $Z = z$ also changes the probability that $\mathrm{X} = x$ [$(P(x|z) \neq P(x)$]. This would leave us with three possible ways in which our three variables could be connected.

1) $Y \rightarrow Z \rightarrow X$: Fixing the value of Y alters the probability that $X = x$ via changing the probability that $Z = z$, which then directly alters the probability that $X = x$ in turn:
2) $Z \rightarrow Y \rightarrow X$: Fixing the value of Z alters the probability that $X = x$ via changing the probability that $Y = y$, which then directly alters the probability of $X = x$ in turn:
3) $Z \rightarrow X \leftarrow Y$: Fixing the values of either Y or Z will directly alter the probability that $X = x$.

Directly connecting Y to X is only correct in case 2 or 3. To disambiguate between these three options, we need to look at whether $\mathrm{P}(x|y) = \mathrm{P}(x|y, z)$. If it does, then once we know that $Y = y$, knowing the value of Z contributes no more information about the value of X—that is to say, knowing $Y = y$ renders the state of X *conditionally independent* of the state of Z. As such, Z can only raise the probability of X via Y, and so we draw our network according to (2).

If, however, $\mathrm{P}(x|y) \neq \mathrm{P}(x|y,z)$, then even once we know the state of Y, knowing the state of Z can still alter the probability that $X = x$. This leaves us with either option (1) or option (3). If $\mathrm{P}(x|z) = \mathrm{P}(x|z,y)$, then knowing the state of Z renders the state of X conditionally independent of the state of Y, and we draw the network according to (1). If, however $\mathrm{P}(x|z) \neq \mathrm{P}(x|z,y)$ then neither Y nor Z screens off X from the other, and so we draw our network according to option (3). By continuing this process with a larger number of variables, we can construct a network like that shown in box 5.1.

A Markov blanket (Pearl, 1988) is a partition on this network that contains all and only those nodes that collectively render the state of some target node X conditionally independent of any other nodes in the network. This means that once the states of this subset of nodes are observed, there will be no further observations that can provide any more information within our model about the probability of *X* being in a particular state—other than to observe the state of *X* directly.

This set amounts to the following: the nodes on which the probability distribution over possible states of *X* is directly dependent (its Markovian parents), those nodes whose state is directly dependent on *X* (its children), and the state of any other nodes on which *X*'s children are also directly dependent (the co-parents)—for a more detailed explanation, see box 5.1.

5.1.1 Bayes Nets in Causal Dress

As described, the immediate advantage of this is in revealing how a complex joint probability distribution may be factored into a number of smaller conditional probability distributions. In this context, a Markov blanket is nothing more than a statistical device—one that allows us to separate out those variables whose state becomes informationally irrelevant with respect to the state of some target variable once we know the state of this smaller subset of nodes. This factorization, as we saw, is exactly the kind of thing that needs to happen in order to implement approximate Bayesian inference via hierarchical predictive processing. Indeed, Markov blankets first appeared in the free energy literature in precisely such a guise (Friston et al., 2007; Friston, 2008).

Still, there is a second more contentious role for Bayesian networks. As developed by Glymour et al. (1993) and later adopted by Pearl (2000), this is their use in causal discovery, allowing us to transform purely observational data concerning the relative frequencies, with which selected variables take particular values into a putative model of the underlying causal structure.

There's a careful line that needs to be walked here. It may be true that no amount of nifty modeling work or nice diagrams will overcome the fact that correlation doesn't *equal* causation. Nevertheless, they are closely related. The growing field of algorithmically driven causal discovery has shown that, when combined with some assumptions, a couple of putative axioms of causation (such as its unidirectionality), and a bit of background knowledge, the purely statistical property of conditional independence may be surprisingly effective in constraining possible hypothesis about where causal relationships may lie.

Box 5.1

Suppose I am trying to figure out what sort of mood my partner, Max, is currently in (*M*)—here modeled as a variable that may be either positive or negative. I cannot observe this directly. Instead, I have to infer it from a limited set of available observations. These include: whether he won his last Rocket League game (*R*), if he's eaten recently (*H*), whether he's grumbling to himself (*N*), and whether he's baking scones (*B*).

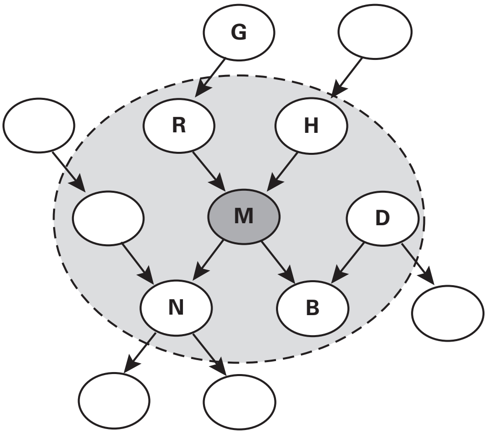


Figure 5.1
A Markov blanket of Max's mood.

I know that if *R* is positive (a victory), then the probability that mood is positive increases, so we can draw an arrow between these nodes, with *M* being downstream of *R*. If *H* is negative, the probability that *M* is positive decreases, so these are similarly connected. I know that a positive *M* also raises the probability of scone-baking, so I can also draw a connection from *M* to *B*. If I want to know the converse, however—that is, how the observation of scone-baking alters the probability distribution over Max's mood—then, as Bayes's rule tells us, I also need to consider any other states that also alter the probability of scone baking, such as whether he has a deadline within the next two days (*D*), which would also increase the probability of baking—irrespective of mood.

Together, these nodes make up the Markov Blanket of M, jointly encapsulating all of the information available in my model about M's current state. Thus, I can take the observations that (1) he won his last Rocket League game ($R=1$), (2) he is baking scones ($B=1$), and (3) he does not have a procrastination-inducing deadline within the next two days ($D=0$), and so infer a high likelihood of a positive mood.

There are other variables in my model, which would also alter my probability distribution over M if I knew their state—such as whether he scored two or more goals in his last game (*G*). These are not a part of the Markov blanket of *M*, however, because a positive value of *G* increases the probability of a positive *M* only by increasing the probability of a victory ($R=1$) Once we know that *R* is positive, the number of goals scored has no further information to offer about the probability of a positive *M*.

In this manner, a Markov blanket renders the target node conditionally independent of the rest of the network.

Returning to our three variables *X*, *Y*, and *Z*, this time with our causal inference hat on, will allow us to more clearly see both the extent and the limitations of a purely observational approach to inferring causation. We saw that looking at whether a third variable renders two variables conditionally independent of each other allowed us to discriminate between three possible graph structures. However, I surreptitiously simplified this problem by sneaking in the assumption that *X* is downstream of both *Y* and *Z*.

From a purely statistical perspective, this is unproblematic. To draw the lines in this direction is not to deny that there is a reciprocal relationship in the other direction. It merely expresses that I am interested in one particular direction of statistical dependence for the purposes of inferring the likely value of *X* given *Y*, rather than vice versa. From a causal inference perspective, however, the choice of direction is a stronger commitment. One motivation behind the use of DAGs in causal modeling is the understanding of causality as a one-way affair. Thus, to draw an arrow from *Y* to *X* is not merely to express which one of two directions of dependence I am interested in. It is to rule out any causal relationship in the opposite direction.

Without this ordering assumption, we would have had not just 3, but 10 possible graphs to consider, all satisfying $\mathrm{P}(x|z) \neq \mathrm{P}(x)$ and $\mathrm{P}(x|y) \neq \mathrm{P}(z)$ (see box 5.2). Further investigation into conditional independence relationships would, in most cases, fail to deliver us one unique graph. Without ordering information, the best we can do is to narrow the possibilities down to one of four sets, three of which contain three further possible causal graphs, all observationally equivalent with respect to a particular statistical relationship between our three variables.

Box 5.2

X and Y are independent given Z: $P(x|z)=P(x|z,y)$, $P(y|z)=P(y|z,x)$
$Y \rightarrow Z \rightarrow X$
$Y \leftarrow Z \leftarrow X$
$Y \leftarrow Z \rightarrow X$

X and Z are independent given Y: $P(x|y)=P(x|y,z)$, $P(z|y)=P(z|y,x)$
$Z \rightarrow Y \rightarrow X$
$Z \leftarrow Y \leftarrow X$
$Z \leftarrow Y \rightarrow X$

Z and Y are independent given X: $P(z|x)=P(z|x,y)$, $P(y|x)=P(y|x,z)$
$Y \rightarrow X \rightarrow Z$
$Y \leftarrow X \leftarrow Z$
$Y \leftarrow X \rightarrow Z$

No conditional independences: $P(x|y) \neq P(x|y,z)$, $P(x|z) \neq P(x|y,z)$,
$Z \rightarrow X \leftarrow Y$

While causality is usually taken to be unidirectional, statistical relationships are not. The occurrence of smoke raises the probability of fire about as much as vice versa, and so the mere fact that A raises the probability of B does not tell us which is the cause, and which the effect. In one sense, this does not matter for the Markov Blanket. The set of nodes included in this is determined by the presence or absence of direct connections between nodes, not their direction. It remains invariant throughout any arrow flips that may legitimately occur while still preserving these same statistical relationships. The set of graphs that preserves the same conditional independence relationships but with different arrow directions is known as the "Markov equivalence class." What may change, however, is which nodes are the parents and which are the children. This will be important when we come to how the FEP uses the Markov blanket construct to factor a system into active and sensory states.

Incorporating time series data resolves some of this directional ambiguity, but there are still problems of unobserved confounders and coincidentally counterbalanced causal chains (Hesslow, 1976). To take Jeffrey's (1969) example of the former, the probability of lightning is increased when the barometer moves to the left, but despite their temporal sequence, the barometer does not cause lightning. No matter how long you observe the two, mere observation of their statistical and temporal relationship will never teach you about the latent common cause of air pressure.

Let's suppose, however, that there are no confounders or counterbalanced causal chains not included in our model. Let's help ourselves to the temporal information required to specify its ordering. Even this would not be enough to deliver a unique model with an unequivocal Markov blanket. Our ability to even ask about possible relations between our variables depends on our having already made a decision about how to parcel our complex molecular world up into the atomic nodes of graph. Some given component of interest may either be broken down into various interacting components or amalgamated into the state of some global variable. For the neuroscientist, should a node correspond to the state of one brain region, one neural cluster, one neuron, one synapse, or one individual neurotransmitter molecule? As with many modeling choices, there is no single answer—beyond saying that it depends on what you're interested in. Bayesian networks, like all models, necessitate idealization and simplification. A more fine-grained model is not necessarily a better one (Cartwright, 2001. Borges, 1998).

Just as the atomic causes of our Bayesian network are a modeling distortion, so too are the tidy little arrows between them (Spohn, 2001; Pearl, 2000). Take a model of conditional dependencies among the membrane potentials of individual neurons. As a Bayesian network, this would represent direct interactions between one neuron spiking and the next. By virtue of the granularity of its variables, said model omits the fact that in order for one neural spike to trigger another, there must be intermediate events of neurotransmitter release and uptake across the synaptic cleft. Interpreting the direct connections of a Bayesian network as direct causal relations, and so understanding the Markov blanket as a boundary made up of the "most proximal" causes and effects, is thus liable to be highly misleading. Anderson (2017) expresses the point nicely in the context of the attempt to locate the "most proximal" boundary of the brain:

> An obvious candidate answer would be that I have access only to the last link in the causal chain; the links prior are increasingly distal. But I do not believe that identifying our access with the cause most proximal to the brain can be made to work, here, because I don't see a way to avoid the path that leads to our access being restricted to the chemicals at the nearest synapse, or the ions at the last gate.

> There is always a cause even "closer" to the brain than the world next to the retina or fingertip. (p. 14)

All of this goes to explain why Pearl (2000) firmly caveats his argument for the utility of Bayesian networks as causal models with an emphasis on the separation between statistical constructs (such as a Markov blanket) and causal facts. A significant body of assumptions and "causal intuition," he notes, are necessary for a network's construction and causal interpretation. This is not too great a concern when our system is made up of nothing more than seasons, sprinklers, rain, and wet grass. In the case of highly complex and poorly understood systems like neural or intracellular networks, however, "causal intuition" will be woefully inadequate to select a single causal model from the innumerable many that will be compatible with the statistical (and temporal) relationships identifiable through observation of our selected variables (Mehler & Kording, 2018).

5.2 Markov Blanket Realism

We have seen that Markov blankets, as they were introduced by Pearl (1988), are a property of a statistical model. Both their constitution and (optional) causal interpretation are dependent on a series of choices and idealizations made in the course of a model's construction. In the work of Friston and coauthors, however, the humble blanket frequently appears not just as a property of our models but of the system being modeled itself. As Friston (2019b) asserts, a Markov Blanket "is not some statistical device by which we come to model the world—it is a necessary attribute of a universe that can be carved into things" (p. 176). As he elaborates with Wanja Wiese, "In the context of the FEP, it is assumed that a Markov blanket is a property of the system itself . . . This version of the concept is therefore a metaphysical notion" (Wiese & Friston, 2021: p. 4).

In the FEP and PP literature, the Markov blanket has not only become more concrete but has been clothed in a variety of unexpected, and almost unrecognizable, guises. Their role in the FEP first gained philosophical attention through the work of Hohwy (2013, 2016), where they become an epistemic boundary that cuts us off from the world beyond our sensory veil—a world that we must nonetheless strive to infer an accurate representation of. Hohwy focuses on brain-bound organisms, using the Markov blanket to draw out sceptical implications for a system that already has a predefined sensorimotor interface. Yet, we do not need the Markov blanket formalism to identify active, internal, and sensory states in the brain, and much early work on the FEP in the brain proceeded without it.

Where talk of Markov blankets has become pivotal to the FEP is in underwriting the application of the active inference formalism beyond the nervous system. As such, we can view their introduction in Friston (2013) as marking a second wave in the development of the FEP, the point at which some of its advocates move from the proposal of a "theory of cortical responses" (Friston, 2005) to a theory of biological systems across spatial and temporal scales, "from single cells to social networks" (Friston 2009a, p. 293).

In this context, the Markov blanket is used firstly to delimit the boundaries of a target system and then, crucially, to divide this boundary up into the sensory, active, and external states that form the free energy equations (Palacios et al. 2020). Here, as Hipólito et al. (2021) describe, the target node is just the internal state, its parents are the sensory states of the bounded system, while its children are the active states:

> In this context, we associate the variable of interest with the internal states of a Markov blanket; which allows us to think of the "parents" of that variable as mediating the influence of external states on internal states (i.e., as sensory states) and of its "children" and the "parents of the children"[2] as mediating the influence of internal states on external states (i.e., as active states). (p. 90)

Note that, as above, despite often introducing the Markov blanket as a "statistical boundary," it is common in the FEP literature to quickly move to causal talk of "influence," "mediation," and explicitly stating that "parent nodes *cause* their children" (p. 90). This causal talk is not accidental. Establishing a one-way direction of influence, not entailed by a merely statistical relationship, is needed to secure a basis for partitioning the system itself into active and sensory states in terms of whether a state directly influences or is directly influenced by the internal state. We saw in the previous section that additional work is required to get from statistical separation to a putative causal boundary and that there is a further gap between the putative boundary of a simple model and "actual causality" in the system itself. These difficulties are largely sidelined in the discussion of Markov blankets within the FEP.

Beyond causal concretization, the Markov Blanket has also gathered a further property, becoming something that is both necessary to preserve the continued existence of a system and something the system preserves in turn. It is not just a description of statistical independence, but rather it is something that "induces" this independence (Ramstead et al., 2019). As Allen and Friston (2018) state, "In short, the very existence of a system depends upon conserving its boundary, known technically as a Markov blanket, so that it remains distinguishable from its environment—into which it would otherwise dissipate" (p. 2475). It is this understanding of the Markov blanket as a real entity, both produced by the organism and causally responsible for preventing that organism's dissipation in turn, that is at issue in the FEP's pretensions to have subsumed the circular relations of self-production and self-distinction described in autopoiesis (Allen & Friston, 2018; Kirchhoff et al, 2018) and to provide "a first principle of living systems" (Friston, 2012).

The task of generalizing the essential features of the autopoietic boundary beyond the molecular membrane of a cell, in order to characterize the autonomy of multicellular lifeforms, has long been a challenge for biodynamic enactivist accounts of life. A possible advantage of the Markov blanket in this respect would be in allowing us to describe the manner in which an autonomous system is separated from its environment, without requiring encapsulation by a continuous physical boundary of the sort conveniently found surrounding the biological cell. Further, as Hesp et al. (2019) propose, such liberality enables the identification of the same free energy minimizing dynamics over a hierarchy of scales—thereby allowing natural selection, social interaction, cultural evolution, development, learning, and planning to all be cast as stages in one grand multigenerational quest for free energy minima.

There is, for instance, and as far as I'm aware, no gelatinous cell-like membrane encasing the nine members of the Bank of England's Monetary Policy Committee. Nonetheless, using the notion of a causal Markov blanket, we can subject their collective behavior to FEP analysis as a stable system that maintains the constancy of an essential variable—the rate of increase in the price of goods—by responding to changes in sensory states, like price indexes, through altering the active states of the base rate and the issuance of government bonds.

You may, at this point, feel that the attribution of vitality to the BoE's Monetary Policy Committee (over and above the limited supply contributed by its individual members) looks less like a success than a suggestion that something has gone horribly wrong with the FEP's analysis of what it is to be a living system. I wouldn't disagree. In the section following this one, we'll see how the definition of active and sensory states in terms of causal Markov blankets can be enlisted to underwrite an even more horrifyingly promiscuous vitalism. The introduction of (causal) Markov blankets to individuate living systems, a task for which they are woefully underqualified is, I will argue, the point at which the FEP starts to lose touch with biological reality.

There are two key issues to unravel in the attempt to understand this piece of the FEP's analysis of life. Firstly, how, if at all, does it make sense to be a "realist" about Markov blankets? Secondly, if a Markov blanket is indeed a property of a real system, and not just of our models of them, then can it do the work that the FEP, in its aspiration to provide a bioenactivist analysis of living systems, has called upon it to do?

The first question has received the majority of attention among critics of the FEP, such as Bruineberg et al. (2022), Menary and Gillet (2020), Beni (2021), and Raja et al. (2021), who accuse Friston and colleagues of reifying aspects of their models. The mistake, as Andrews (2021) puts it, of confusing the "math for the territory." Still, Markov blanket realism doesn't have to be a mistake. All the above authors argue is that Friston and colleagues have not supplied the metaphysical premises required for their metaphysical conclusions.

That doesn't mean no such premises are available. Nor would it require a commitment to mathematical Platonism as Beni (2021), Menary and Gillet (2021), and Bruineberg et al. (2022) suggest. There is a positive position one could take on the structure of reality and the metaphysics of causation such that it would be plausible to take Markov blankets as real entities, not just modeling constructions. This position, and some potential motivations for it, is discussed more extensively in the appendix, but in brief, it involves two aspects: probabalistic graph realism and the statistical reduction of causation, perhaps supplemented by a Penelope Maddy–style naturalized mathematical realism about sets as having an existence over and above that of their parts (Maddy, 1990, 1997).

Probabilistic graph realism means taking reality itself to literally have the structure of a Bayesian network, being made up of independent nodes whose state is determined exclusively by local interactions with other nodes, which respect the Markov condition—such that the state of each unit is independent of the state of its non-descendants, conditioned on its parents.

Positions related to this have been defended by philosophers such as David Lewis (1994), David Papineau (1992, 2022), and Hartry Field (2003).

This view is often, but not always, combined with a statistical reduction of causation, which takes the statistical relationships described in a Bayesian network to be all there is to causal relationships. On such a view, the reason that correlation does not equal causation is not that they are different things entirely, but just because we haven't gathered *enough* correlational information to uniquely determine causal structure. Such a view has been defended by Spohn (2001), Reichenbach (1956), Good (1959), Suppes (1970), and Papineau (1992) (and for reviews, see Salmon [1980] and Weslake [2006]), though there is disagreement on whether this accounts for causation as a feature of reality or just an inescapable feature of our conceptualization of it. If causation does reduce to statistical relationships in this way, then the Markov blankets of the ultimate graph of reality become *causal* boundaries, in the way Friston and colleagues often assume.

One further commitment is required for these real Markov blankets to be something that can demarcate an organism, as Friston (2013) and Allen and Friston (2018) suggest. This is the requirement of what I'll call a *stable* Markov blanket, namely, that the structure of the graph corresponding to the organism—determined by the patterns of interaction between its parts—is stable enough such that there is a fixed set of components that renders it conditionally independent of its exterior and this set endures for the duration of the organism's existence.

As Friston (2013) puts it,

> A candle flame cannot possess a Markov blanket, because any pattern of molecular interactions is destroyed almost instantaneously by the flux of gas molecules from its surface. Meaning we cannot identify a consistent set of blanket states rendering some internal states independent from other state. (p. 2)

So, all that positing a real Markov blanket involves is the claim that a system decomposes into independent units (accomplished by some means prior to the FEP, by methods the framework itself does not specify) and arguing that if we accept objective statistical dependencies as an adequate reduction of causal relationships, along with the principle of locality, then the immediate surroundings of any one of those units will literally have the Markov property of inducing a conditional independence between what is inside this boundary and what is outside of it. To claim that Markov blankets are real things is to make a general claim about the structure of the causal universe. In doing so, the FEP has neither identified a new and interesting entity in the world nor discovered a principled basis for carving the world into things. This latter result depends on the necessarily prior task of telling us what the absolute units of the ultimate graph are.

Whether *we* should care about these metaphysical questions depends on the second question, regarding whether Markov blankets are actually of any importance in a theory of living systems. In the next chapter, I will argue that Markov blankets are of no help at all in helping the FEP to formalize and "subsume autopoiesis" (Korbak, 2021, p. 2747) or to "supersede or absorb classical (i.e., autopoietic) formulations of enactivism" (Ramstead et al. 2021, p. 59). It is hard to see why one would think they could. Instead, the FEP's claim to perform this role, as it turns out, depends on an implicit assumption about the cyclical structure of a system's causal graph. This provides an alternative demarcation of said system, prior to our factoring it up with a Markov blanket. While this particular graph structure, as I will describe, does bear some similarities to early definitions of autonomy, it bears no obvious relation to contemporary accounts in terms of a closed network of precarious processes.

There is a second reason to be less concerned with the reality of Markov blankets than with their relevance for living systems, as I will describe in chapters 8 and 9. This is the observation that even if some parts or aspects of reality do have the structure of a statistical-causal graph, an organism is about the last place we would expect to find the stability of interactions needed to pick out a *stable* Markov blanket as persisting and defining the existence of this system over time.

Taking Stock

Before we move on to the evaluation of the free energy principle (FEP) as a formalization of bioenactivism, a quick refresher of the key concepts introduced so far.

The **free energy *principle*** is the claim that every system of a certain type (microcircuits, brains, organisms or literally every "thing," depending on the claimed scope) must minimize the free energy of its constituent parts in order to continue to exist. **Free energy** is a function of two things. One: **divergence**, the difference between a simplified **recognition model** that a system encodes and the actual statistical properties of the process that generates its evidence, called the **generative model**. And two, **surprisal**: the unlikeliness of a particular state relative to this generative model.

It turns out that minimizing the first thing matters only insofar as it positions a system to minimize the second. Most discussion of the free energy principle as it applies to nonneural systems thus disregards any encoded recognition model to focus on the "true" generative model and the minimization of surprisal.

Taking the minimization of surprisal to be necessary and sufficient for self-preservation rests on a definition of ongoing existence in terms of maintaining a **steady state** (Friston & Mathys, 2016). "Steady state," crucially, does not require that the state of the system never changes, only that any changes will preserve the same statistical properties such that the probability distribution over potential states of the system at any randomly selected point in time is stationary over the duration of the system's existence.[1] This is compatible with the presence of symmetric and stochastic fluctuations, with a consistent amplitude, in the system's state, but these must be countered by a return to more likely states, described by the **dissipative flow**. It is also compatible with, though does not require, the possibility of cycles between the same subset of equally likely states, described by a **solenoidal flow**. Further, the FEP also requires that this stationary probability distribution has **low entropy** such that the repertoire of states said system is likely to be in is not only constant but also relatively small.

The argument for presenting these components as jointly definitive of existence was as follows. On one hand, stationarity alone would be satisfied if a wide range of states remained equiprobable over time—but in such a case there would be no distinctively characteristic states by which the system could be re-identified. On the other hand, if the system was likely to be in only a small region of states at each time period, *but* this region was constantly changing, then we would have no way to reidentify it as the same system persisting over time. The paradigmatic example given of steady-state dynamics is biological homeostasis, though, as we saw in chapter 4, it can also be extended to describe the stability of non-biological systems.

The connection between the minimization of free energy and the maintenance of a non-equilibrium steady state is supplied by **active inference**. This extends the idea of variational free energy minimization, a means to adjust an approximate recognition model to better fit one's evidence, by adding in the option of acting so as to change this evidence instead. Such a system can thus minimize free energy in two ways: (1) by changes in the internal states that encode its recognition model, thereby "learning" what its characteristic sensory inputs are—as described by the generative model of said system, and (2) by acting to minimize the surprisal of its sensory evidence, thereby countering dispersal away from these characteristic states. This rests on the idea that in reducing the surprisal of sensory states, a system implicitly reduces the surprisal of the distal states that characterize it as the kind of system it is.

Importantly, however, the "inferential" interpretation of this rests on the claim that the system actually encodes this "recognition model." Neither the claim that a system "embodies" or "entails" a generative model nor the assertion that it "minimizes surprisal" implies that anything inferential is going on. All this means is that the system has steady-state dynamics with the tendency to return to the same set of states when perturbed, such that we could describe it *by* the stationary joint probability distribution of a generative model.

Still, when there is no recognition model, and no divergence between one and the true generative model, then, mathematically, free energy does reduce to surprisal. As such, the FEP takes something's merely being a steady-state system, describable by a generative model, to suffice for describing it as actually "using" this model itself, in order to infer and to minimize its "free energy."

To do this, the FEP needs to decompose the system's constituent states into internal, sensory, external, and active variables. And it is this decomposition that is supposed to be accomplished by means of the **Markov blanket** formalism just discussed.

All of these constructs together make up the free energy *framework*.

The upshot of this is the claim that the dynamics of a system at steady state can be described as formally analogous to Bayesian inference. But is a steady-state system with a Markov blanket enough to formalize anything like autopoiesis, autonomy, operational closure, or organizational closure? And, even if it is not, does it provide an alternative definition of a living organism?

6 From Markov-Blanketed Steady-State Systems to Sensorimotor Cycles

If we are willing to allow some pretty hefty assumptions about the nature of reality, then any stable entity we pick out will have a Markov blanket composed of a further set of elements that suffice to make its state conditionally independent of everything else. Does this, combined with the requirement that a system is at a surprisal-minimizing steady state, give us everything we need to start talking about active inference?

Friston (2013, 2019b) certainly claims as much, declaring that an ergodic (or steady-state) system and a (stable) Markov blanket are all that is needed to get active inference off the ground and to interpret internal states as changing so as to minimize variational free energy with respect to a probabilistic model. “With this existential dyad,” he claims, “everything of interest about life and the universe can be derived, from biotic self-organisation through predictive processing to the detailed microcircuitry of our brains” (p. 176).

We need at least these two requirements as neither a *stable* Markov blanket nor steady-state dynamics entails the other. Under probabalistic graph realism, a charging battery has a Markov blanket, but over that duration, its state of charge is increasing, not steady. Similarly, various stable chemical reactions might be describable by a fixed probability distribution over some collective property, like average concentration, but due to a constant flux of materials, there would not be sufficient stability of interactions between these parts to establish a stable Markov blanket.

So, a particular system will be a candidate for analysis in terms of active inference only if it has both (a) stability in terms of the changing states of each of its parts—which gives us the steady-state distribution or generative model—and (b) stability of interactions between those parts—which gives us a *stable* Markov blanket.

Still, there are a heck of a lot of stable physical systems with a boundary more persistent than a candle flame that look nothing at all like a living cell. Without its chocolate casing, a cream egg would disintegrate. Getting at the cream requires breaking through this shell. The air pressure from the egg's surroundings may be cast as a "sensory" state and the countervailing force of the cream pushing against its chocolate cage may be cast as an "active" one that counteracts this pressure to maintain a stable eggy shape. But for all that the chocolate egg renders its gooey innards conditionally independent from the world outside; a cream egg is neither sentient, autopoietic, nor autonomous.

How then, are we to understand claims like the following? "Life—or biological self-organization—is an inevitable and emergent property of any (ergodic) random dynamical system that possesses a Markov blanket" (Friston, 2013, p. 1).

And such Markov-blanketed systems are

> autopoietic: because active states change—but are not changed by—hidden states, they will appear to place an upper (free energy) bound on the dispersion (entropy) of biological states. This homoeostasis is informed by internal states, which means that active states will appear to maintain the structural and functional integrity of biological states. (p. 5)

Similarly, Kirchhoff et al. concur (2018) that "any Markov blanketed system will embody recurrent processes of autopoietic self-generation, which—as long as the system exists—enforces a difference between a living system and everything else" (p. 6).

This point is made more expansively by Allen & Friston (2018) as follows:

> For example, a cell persists in virtue of its ability to create and maintain a boundary (cell-surface), through which it interacts with the environment, thereby maintaining the integrity of the boundary. It is this autopoiesis, or self-creation, which enables the system to limit the possible states it visits, and thus to survive (Varela et al., 1974). The FEP recasts this as a kind of self-fulfilling prophecy, in which an organism itself constitutes, in the generative sense, a belief that it will prevail within certain embodied and environmental conditions. In short, the very existence of a system depends upon conserving its boundary, known technically as a Markov blanket, so that it remains distinguishable from its environment—into which it would otherwise dissipate. (p. 2473)

Such statements rarely acknowledge the distinction emphasized by both Maturana and Varela (1980) between autopoiesis as a recurrent process of metabolic self-assembly, as opposed to the more general concept of autonomy that attempts to generalize this logic of self-production beyond the molecular interactions of a single cell. In as much as autopoiesis is constrained to a specific level of chemical interactions, no purely statistical generalization could capture it. Elsewhere, however, free energy theorists target the more general notion of autonomy, which, despite general agreement to the contrary (Thompson and Di Paolo, 2014; Bich and Arnellos, 2012; Di Paolo et al., 2017), they take to be adequately expressed in the concept of operational closure.

It is, Ramstead et al. (2021) claim, "fairly straightforward to establish that the Markov blanket formalism provides a statistical formulation of operational closure." (p. S55). Given that the kind of circular relationships of input to output that define operational closure have no part in the concept of a Markov blanket, the idea that they serve as a formalization of the former concept appears far from straightforward to me.

Sadly, the paper by Kirchhoff et al. (2018) that Ramstead et al. cite as performing this "straightforward" establishment does nothing of the kind. The main justification provided for relating the Markov blanket to operational closure is by way of reference to what Varela calls "the intriguing paradox" of autonomy in how it requires that living systems are closed off and distinguished from their environment, while at the same time completely dependent on remaining coupled with that external environment for their ongoing existence. Kirchhoff et al.'s suggestion is that the conditional independence of internal states from external states describes this closure, for once we have determined the blanket states, we are closed off to the possibility of gaining any further information that might reduce our uncertainty in predicting internal states. At the same time, internal states are still open to influence from external ones, *via the blanket*. The Markov-blanketed system thus exhibits a balance of closure to information (conditioned on the blanket) and openness to causal influence (via the blanket).

As an analysis of autonomy, this is unconvincing. Being an autonomous system, as we will see in more detail in chapters 7–11, is not just a matter of exhibiting closure to some things and openness to others, but a matter of being closed to the *right* things. A formalization of autonomy is not provided by describing any old mixture of openness and closure but rather by identifying precisely what particular things an autonomous system must be open or closed to. Nonetheless, Kirchhoff et al. move on from this to claim that:

> This teleological (Bayesian) interpretation of dynamical behaviour in terms of optimization allows us to think about any system that possesses a Markov blanket as some rudimentary (or possibly sophisticated) 'agent' that is optimizing something; namely, the evidence for its own existence. This means we can regard the internal states (and their Markov blanket) as, in some sense, autonomous. (2018, p. 2)

We can also think of a wobble doll as believing that remaining upright is its highest calling, and actively striving to achieve this. Or, as Kirchhoff et al. acknowledge, we might consider a pendulum as engaging in active inference. But just because we can doesn't mean that we should. Surely it takes more to legitimate agential and inferentialist language than just stability in the face of perturbations.

Elsewhere, Friston, et al (2020) instead target the notion of "sentience," pointing to the ability to divide a Markov-blanketed system into parents and children of the target node, where the parents of some node are those upstream that directly affect it and children are those downstream that are directly affected by it.[1] The idea of a sensory state, they claim, is captured in nodes whose parents are external states, and whose children are internal states—and vice versa for active states.[2]

This is no more compelling as an analysis of sentience than it is of autopoiesis. If these dependency relations were all there were to an active or sensory state and if, as Friston et al. (2020) advocate, being "responsive to sensory impressions" is all there is to sentience, it looks like we are ascribing capacities of sensation and action to everything that has—or at least can be modeled as having—a stable Markov blanket over some duration.

Friston et al. (2020) attempt to elide the panpsychist implications of this by noting that their attribution of sentience "is not used in the philosophy of mind sense; namely, the capacity to perceive or experience subjectively, i.e., phenomenal consciousness, or having 'qualia'. Sentience here, simply implies the existence of a non-empty subset of systemic states; namely, sensory states" (p. 3). This may be enough to avoid attributing a rich inner life to everything with a Markov blanket, but one still wonders what possible justification could then be offered for using the terms like "sentience" and "sensory" here. Taken in the most minimal sense of implying the capacities of sensation and action, attributing sentience on the basis of a Markov blanket alone is still going to lead to a mathematically motivated animism that considers steam engines, pendulums, plants, and people as all sentient systems alike.

There must then be still more baked into what Friston and colleagues mean by a "Markov blanket" than the properties of stability and physical realization.

6.1 The Missing Cycle

Now that we are more familiar with Markov blankets, claims of such a straightforward connection to autopoiesis, autonomy, or sentience, and so to the emergence of life, agency, and consciousness, should immediately appear rather suspect. I imagine that the fact that such proclamations have remained unchecked for so long is due to the way Markov blankets are typically explained in the FEP literature.

Firstly, there is usually little offered by way of introduction other than the claim that they are a statistical partition, which then quickly morphs into causal talk of "influence" when introducing the division into descendants and ancestors of a target node.[3] The interpretations of the Markov blanket as a mathematical object, or as a physical boundary, are treated interchangeably, and there is typically little, if any, mention of either (a) the various assumptions of the modeling framework of causal graphs, which underpin the initial partition, or of (b) the metaphysical premises required to secure its causal interpretation and physical concretization.

Secondly, the paradigmatic example chosen to illustrate a Markov blanket is almost always the cellular membrane (Friston, 2013; Palacios et al., 2020). While the cellular membrane is indeed a physical boundary surrounding the cell and would thus be a Markov blanket in the supposed all-encompassing graph of reality, the properties of "self-generation" and "self-preservation" that Friston and colleagues go on to attribute to it are held, not in virtue of this, but only in virtue of its also being specifically an *autopoietic* boundary—something that the vast majority of physical boundaries and Markov blankets are not. This is rather like taking the true statements that "Amy is my friend" and "Like all my friends, Amy is a featherless biped" and then proceeding to suggest that I have successfully analyzed the concept of friendship.

So far, I've engaged in some metaphysical speculation to justify interpreting a Markov blanket as a physically instantiated causal boundary and supplemented this with an additional requirement for the stability of said boundary. I have also pointed out that the FEP is concerned only with the Markov blankets of steady-state system, and as such it requires not only stability of dependencies between parts but also stability in the tendencies of these parts to occupy a particular subset of states. Yet even this pumped-up “stable Markov blanket plus steady state” description remains incapable of supporting the claims that are made of the “Markov blanket formalism” in the FEP literature.

Importantly, however, in as much as the steady-state Markov blanket fails to entail the circular dependence involved in autopoiesis or autonomy, it also fails to capture the kind of perception-action cycles described in active inference. As presented in the context of a brain, active inference was not just about a system that remains at steady state, and in which some parts are statistically shielded from others, but also one in which these parts exert a distinctively circular pattern of influence on each other. In particular: it described a system where a discrepancy in the sensory part of the system, relative to the internal part (interpreted as encoding a predictive model of these sensory states) leads to a change in action, which changes the external environment so as to alter the sensory state and reduce this discrepancy. How is this circular pattern of influence derivable from a Markov blanket of either an ergodic or steady-state system?

The answer, and I’m sure you can imagine how frustrating this would be had you (hypothetically) spent months trying to understand how active inference can be derived from these explicitly declared and oft-repeated prerequisites, is that it can’t.

Markov blankets, as discussed previously, are traditionally used in the context of a directed *acyclic* graph (DAG). In the philosophical literature on the FEP, this is typically the format in which Markov blankets are depicted. The problem is that the acyclicity of this type of graph specifically prevents depicting the kind of circular connection needed for active states to have a reciprocal influence on the sensory nodes that are their (indirect) ancestors.

Nonetheless, acyclicity was not part of the requirements laid out for realism about probabilistic graphs. It is indeed possible to have a graph with cycles that are factorizable by Markov blankets, though it is less straightforward than in the acyclic case. Accordingly, there is a second, quite different, graph also presented as depicting the structure of an active inferrer that appears in Friston and colleagues’ discussions of the FEP (see figure 6.1) typically found rakishly imposed over a picture of a brain. Unlike the acyclic structure of figure 6.1a, this second diagram, figure 6.1b, explicitly depicts the kind of cycle between external, internal, sensory, and active

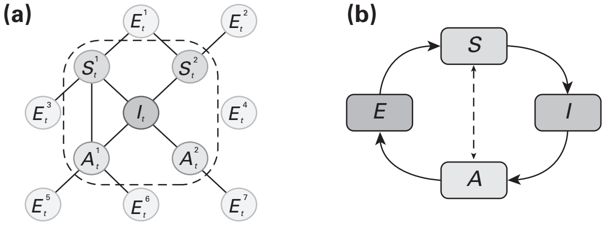


Figure 6.1
Depicting a Markov blanket in (a) the standard acyclic graph and (b) the cyclic graph that is typically presumed in work on the free energy framework—from Rosas et al. (2020).

states (ESIA) that would be expected in the perception-action loop of active inference.

The first diagram is derivable from the assumption that something is divisible into parts whose interactions respect the causal Markov condition, as discussed in the previous section. This "ESIA cycle," as depicted in the second diagram, is not. Rather it depends on the ability to partition our overall steady-state system (x), described by a stochastic differential equation (equation 4) into a set of coupled equations describing four sets of variables, external (E), sensory (S), internal (I), and active (A) and how they influence each other—as shown in equations 4 and 5:

$$\dot{x} = f(x) + \omega \tag{4}$$

and $f(x) =$

$$\dot{\mathrm{E}} = f_E(E, A, S) + \omega_{\mathrm{E}}$$
$$\dot{\mathrm{S}} = f_S(E, A, S) + \omega_{\mathrm{S}}$$
$$\dot{\mathrm{I}} = f_I(I, A, S) + \omega_{\mathrm{I}}$$
$$\dot{\mathrm{A}} = f_A(I, A, S) + \omega_{\mathrm{A}} \tag{5}$$

Said equations specify how each set of variables changes as a deterministic function of the state of a subset of other variables, plus a noise term (ω) representing the stochastic fluctuations that the surprisal-minimizing tendency of our system is stipulated to dissipate in order to remain at its steady state. Where the Markov blanket induced a requirement for the stability of statistical dependencies between the parts of a system, the ESIA-cycle equations describe one possible specific flow of interactions that might meet this requirement, namely a flow in which internal states are a function only of sensory, active, and other internal states, not of external states directly (though see Rosas et al. (2020), Biehl et al. (2021), and Aguilera et al. (2022) for a more technical criticism of the relation between the two).

So, it is the ESIA-cycle equations of figure 6.1b, rather than the Markov blanket itself, that specifies the cyclical flow of the sensorimotor loop involved in active inference. These equations have received limited attention in philosophical discussions of the FEP, however, where the claim that the Markov blanket is what determines the boundaries of the system of interest has largely been taken as read (Hohwy, 2017; Clark, 2017; Kirchhoff & Kiverstein, 2021). If a diagram showing the cycle of figure 6.1b does appear, it is often treated as a straightforward depiction of a Markov blanket and presented without any explanation of these equations on which it is based (Hohwy, 2017; Kirchhoff & Kiverstein, 2021).

The move from the presence of a Markov blanket to the identification of an ESIA cycle is not a straightforward derivation. However, there is a simple, though rather disingenuous, trick that can turn the former into the latter. First, we take all of the external variables E_1, E_2, . . . E_7 and redescribe them as a single coarse-grained macro-variable E^*, whose overall state depends on the combined state of all those individual external variables. We might also do the same for our sensory and active states, treating the former as a single macro-variable S*, describing the entire state of the organism's sensory periphery, and the latter as another macro-variable A* determined by the entire state of motor outputs. Then we note that the single E^* variable is simultaneously the only external variable that active states can influence, and the only external variable that can influence sensory states.

Voila! We have a cycle from active to sensory states via the closed loop of E^*.

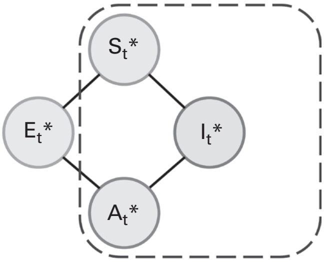


Figure 6.2

A Markov blanket in which all the external variables E1–E7 in figure 6.1a have been condensed into one macro variable E^*—based on Rosas et al. (2020).

This reflects nothing about the nature of any system we might identify a Markov blanket around. The ability to rustle a cycle up out of anything only highlights the point made in the previous chapter: probabilistic graphs tell us nothing informative about a system unless we already have principled means to divide this system up into the fixed units or variables that make up such a graph. If the parcelling-up process is completely arbitrary, then so too are the graphs that we construct.

We might, for instance, describe the color of my jumper and the exact position of the balls in The National Lottery's *Exacalibur* machine as composing a single macroscopic variable: *Jumpscalibur*. It is technically true that the state of the combined variable, Jumpscalibur, this Saturday evening determines whether or not I will be a millionaire on Sunday morning. It is also true that whether or not I wear my lucky orange jumper will change the state of Jumpscalibur. Still, only the most wishful of thinkers could believe that the color of my jumper might somehow be responsible for manifesting millions of pounds into my pocket. Permitting the arbitrary aggregation of variables allows us to create statistical relationships between unconnected events. When these relationships are interpreted causally, this produces the appearance of a causal relationship where no such relationship exists.

So, rather than attempting to justify the derivation of an ESIA cycle from the presence of a Markov blanket, it makes more sense to take the assertion that the interactions between parts of a system have the structure of an ESIA cycle to be an independent and prior assertion. Based on knowing this structure, we may then derive a Markov blanket for said system. There is an interesting parallel here to the way in which the idea that a system is identical with a generative model turned out to be dependent on the prior assertion that the system had steady-state dynamics. In both cases, our proposed probabilistic first principles turn out to be the redescription of implicit dynamical postulates.

By establishing a cyclical flow, the ESIA equations specify a more restrictive requirement for active inference—over and above a steady state and a Markov blanket—namely, for a pattern of feedback from active to sensory variables, via external ones. This cycle, as Rosas et al. (2020) note, looks closer to the kind of thing that we think of as a sensorimotor loop and makes the identification of constituent variables as "sensory" or "active" a slightly less trivial matter than when it is done in terms of Markovian ancestry alone.

Beyond this, the "ESIA-cycle" equations also tell us which external variables should be included in the overall model of the system-in-its-environment, and which can be excluded. The Markov blanket itself could not do this because, as far as a Markov blanket is concerned, one external state is just the same as any other no matter how far away in the causal chain. All are equally irrelevant to predicting internal states once the states of the blanket nodes are established. In figure 6.1a, the inclusion, or exclusion, of an external variable looks like an arbitrary choice. Once we have the cycle defined by equations 4–5, and depicted in figure 6.1b, we can see that the only Markov-external variables that are relevant from the perspective of the free energy framework are those that are part of this cycle—namely those Markov-external variables whose state is both a function of active variables and a partial determinant of the state of sensory variables.

This helps make sense of the FEP's rather implausible requirement that it is not only the internal, active, and sensory states of a system but also its external variables that must be at steady state. If this were not the case, if internal states converged to a stable regime while external ones were constantly changing, then the former could not be regarded (even in the weakest correlational terms) as entailing a model of the latter. As Millidge et al. (2021a) note, if steady state were a requirement taken to apply to the universe as a whole, it would (theories of eternal recurrence aside) be quite obviously false. The ESIA-cycle formulation allows us to restrict the requirement of steady state to the Markov-blanketed system's niche, or umwelt, now defined as all and only those "external" variables that are part of said cycle. In this sense, the ESIA cycle captures the cyclical nature of active inference, introduced in section 3.2, under which the stability of external states is a by-product of the Markovian system's activities in working to maintain its observations and internal variables at steady state.

What then of the extra-external world, outside this ESIA-defined niche? As described in section 3.5, the active inference framework attempts to distinguish between this external generative process and the agent's generative model. As we saw there, such a distinction becomes difficult to maintain once we recognize the cyclical relationship between an agent's observations and actions. In the ESIA-cycle equations, however, we can see that there is some residual role for the influence of causes originating outside of the active inference agent's circular dynamics. Specifically, these causes are reduced entirely to random fluctuations—to the normally distributed, independent, additive noise terms (*w*) that are stipulated as sufficient to capture the only kind of influence that the ESIA-external world can have on the steady-state system (Millidge et al., 2021a).

This idea that we can adequately describe a system and its local niche in terms of this simple steady-state cycle by condensing all external influences on it to uniform noise, is somewhat more tenable than the requirement that the entire universe be in a stable dynamical regime—though not much. I have already hinted at the implausibility of suggesting that the entire life cycle of an organism can be modeled by the convergence to and maintenance of a single stationary probability distribution and I will discuss this further in chapter 9.

The free energy formulation's dependence on this steady state may, as Millidge, Seth, and Buckley suggest, make more sense as a locally valid approximation. Still, the fact that our environment is so volatile and changeable, not only in its temporary state but also in its ongoing dynamical tendencies, is precisely what makes biotic systems' ability to maintain their homeostatic stability so interesting. To *begin* by modeling the system's local surroundings as being at steady state would seem to extract much of the interest from this problem at the get-go. As Millidge et al. (2021a) put it:

> This is the real question which is what the FEP tries (at least in its intuitive sales pitch) to answer—how can I maintain an internal steady state against an environment which is *not* at steady state. By assuming that the external states are also at the steady state, it may be that the FEP is, in some sense, answering the wrong question and is, in the process, assuming away the true difficulty in answering the right one. (p. 34)

We've seen then that the Markov blanket and the joint probability distribution of the generative model are not so much explanations of how a system can preserve its homeostatic stability in a changing world, but rather redescriptions of a system in which this problem has already been solved. That system is described by the stochastic differential equations 4 and 5, and such equations are precursors to a FEP-style description, rather than being derivable from free energy minimization as a first principle.

Still, the system described by those equations, shown in figure 6.1b, certainly appears closer to being a candidate of operational closure, or a perception-action cycle, than figure 6.1a.[4] If we say a sensory or active state is determined not by its Markovian ancestry but rather by its membership of such a cycle, then we have a slightly more restrictive definition of these states than one that depended on conditional (in)dependence alone. But, before evaluating the suitability of this formulation to capture the particular kind of closure that distinguishes living systems, we need to look a little more closely at how the system defined by these ESIA-cycle equations meshes with the claims that have been made about the role of the Markov blanket.

6.2 Extensional Ambiguity

With the ESIA-cycle equations and the Markov blanket both in hand, we now have two different partitions and two different sorts of "externality" involved in the definition of a single system: things that are external to the Markov blanket but internal to the ESIA cycle and things that are external to both. As such, as Raja et al. (2021) note, Friston's "existential dyad" is ambiguous on whether it is the overall steady-state ESIA cycle, or only the Markov-blanketed subset of this, that determines the boundaries of the system that the free energy principle is supposed to define.

In some places, the Markov blanket is described as what differentiates "between the system and its environment—those states that constitute or are intrinsic to the system and those that are not" (Ramstead et al. 2018, p. 3) and as providing "a statistical formulation of operational closure" (Ramstead et al. 2021, p. 55). Yet, the steady-state equation, which ranges over the entire ESIA cycle, is also described as capturing "the phenotype" of an organism (Ramstead et al. 2020) or even, as Friston (2019) more broadly declares, something that can capture the entire concept of "thingness." This seems at one remove to consider something as external to our organism (or other system of interest) and at another to consider it as part of the phenotypic states that define it.

We have already discussed the inadequacy of the Markov blanket to formalize anything like self-organization or autonomy, to identify a sensorimotor interface, or, indeed, to carve out anything particularly unique at all. The natural move then would be to turn instead to the ESIA cycle, which we will do in the next section. But if we take this to provide both the demarcation of our system and its decomposition into active, sensory, internal, and external states, then there is a residual question: What is it that the Markov blanket demarcates?

As with the "simple imperative" of surprisal minimization, I think we can gain some insight from Ashby's work on self-organization here, specifically a procedure he suggests for describing the appearance of an intelligent system. (Note that "equilibrium" is used in the looser sense of stability or stationarity of dynamics here, and it is compatible with a system being at non-equilibrium steady state or "dynamic equilibrium.")

> Take a dynamic system whose laws are unchanging and single-valued, and whose size is so large that after it has gone to an equilibrium that involves only a small fraction of its total states, this small fraction is still large enough to allow room for a good deal of change and behavior. Let it go on for a long enough time to get to such an equilibrium. Then examine the equilibrium in detail. You will find that the states or forms now in being are peculiarly able to survive against the changes induced by the laws. Split the equilibrium in two, call one part "organism" and the other part "environment": you will find that this "organism" is peculiarly able to survive against the disturbances from this "environment." (1962, p. 120).

The key idea, here and elsewhere in Ashby's discussions of self-organization, is that the appearance of any intrinsic volition is illusory. What we call "self-organization" is really driven by the coupling between our selected system and another one. We might artificially decompose these and treat each as an individual system, but neither has intrinsic stable dynamics that are independent of their interaction. The Markov blanket is a means of making a similar division, but in describing it as the boundary of an organism, the free energy framework has unwisely elected to drop the scare quotes that are essential to reminding us that this distinction between the two is a choice imposed by an external observer.

The problem with taking the Markov blanket as the boundary of an organism, rather than a pragmatic division, can be made more concrete in the case of a system that is both enbrained and embodied. Here, this Markovian partition could cut around the nervous system—in which case bodily variables (e.g., core temperature) and environmental ones (e.g., air temperature) take on the same equal status as undifferentiated "external" variables. This makes sense in that both are part of what we may want to describe the nervous system as modeling and regulating, but it erases the distinction between the body and the environment, allowing no special import to be given to the management of the former. If, instead, we take the relevant Markov blanket to be that which divides the whole body from the environment, then the internal relationship between neural-regulator and regulated body is washed out in favor of describing the entire body, en masse, as a regulative model of its surroundings.

This is not just the point that I have already drummed to death about the variety of scales at which Markov blankets can be identified. As Ramstead et al. (2019) and Sims (2021) suggest, this could naturally be taken as analogous to the multiscale nature of organisms composed of organs, organelles, and cells. The bigger issue here is that at neither the scale of the brain, nor of the whole body, does the Markov boundary adequately parcel up the system we are interested in as a brain-equipped organism.

Ashby seems to take the mere appearance of one system's adapting to another as sufficient to ground an account of intelligence in terms of the complexification of this process. As a consequence of choosing to treat the differences between inanimate matter, organisms, or intelligent systems as a gradient, and homeostasis as a general property of stable systems, his position is undisturbed by the concern that this decomposition into "organism" and "environment" is a modeling heuristic that can be applied to anything from a Watt governor and a steam engine or a pair of coupled pendulums.

In the context of this Ashbyian approach to self-organization, the question of whether a particular Markov blanket truly demarcates an organism would reduce to a matter of whether the active inference analysis that a Markov blanket facilitates grants sufficient explanatory power to pay its way in our best scientific theory of the system. This is a much weaker view of what the free energy principle delivers than has often been claimed—through Friston et al. (2020) do entertain such an instrumentalist stance with regard to the attribution of sentience, specifically as a means to avoid the panpsychist consequences of a Markov blanket–based demarcation of the mind.

Extending this instrumentalism further, to the demarcation of life itself, seems a more radical (and ethically unpalatable) position. Nonetheless, if the FEP's advocates are prepared to treat the distinction between organism, mechanism, and other material objects as a matter of degree—the demarcation between them as something dictated by the interest of an observer—then this would at least render the free energy principle's more philosophical claims consistent with their Ashbyian underpinnings.[5]

Continuing in this Ashbyian vein then, a plausible alternative to treating the Markov blanket as something that demarcates a living system might run as follows: any organism (a steady-state ESIA cycle, demarcated by non-Markovian means) will be divisible into cognitive-regulat*or* and bodily-regulat*ed* parts, and there will be a Markov blanket between these. Markov blankets, being exceedingly trivial if they are to be considered real features of our world at all, cannot suffice to exclusively identify this demarcation, but the fact that said division is also a Markov blanket may allow us to say some interesting things about the relationship between these parts.

For instance, we can note that as each Markov-blanket-individuated set of parts: internal, external, active, and sensory, has steady-state dynamics derived from those of the overall organism, so each will each tend to spend most of its time in the same characteristic state (or cycle of states if the overall system also has a solenoidal flow), returning to these when perturbed by internal or external fluctuations. If we describe the stable dynamics for each component in terms of a probability distribution over most likely states, then we will be able to interpret each component's return to its characteristic state (or cycle) when perturbed as minimizing surprisal relative to this probability distribution. As each component is doing this, there will be low KL-divergence between the respective probability distributions of all these components. Each time one component returns to its lowest surprisal state when perturbed, it may be interpreted as "preserving" a stable probability distribution and minimizing the KL-divergence between its average state and that of other parts of the overall steady-state system.

Still, all that "inferential" description reflects is the covariational dynamics that are inevitable among the parts of any system that has the kind of overall stability the FEP requires. "Inference" here is not a means to obtaining this stability, but rather a redescription of it, and there is nothing especially intentional, agential, or representational about the kind of systems that admit such redescription. As argued in section 3.5, to straightforwardly take KL-divergence as a measure of misrepresentation, and its ongoing reduction as evidence of something's performing a representational function, would be to reduce representation to covariation and license the attribution of representational relationships between the parts of any stable system.

As Raja et al. (2021) note, it is not immediately clear that we gain anything by describing this relationship in Bayesian vernacular that we did not have access to from a description in terms of the dynamics of feedback control loops between coupled subsystems. Markov blankets, as they put it, appear to be nothing more than a "trick" of exploiting the general divisibility of (most) systems into parts whose interactions respect the Markov condition, to redescribe their relationship as the inference of one to the other. Whether you take this to be reflective of fundamental metaphysical truth, or merely a feature of a particular modeling convention, either way, it does not reflect anything uniquely teleological, cognitive, agential, or animate about the particular systems that we chose to execute the procedure upon.

As with mere covariation, this general property may gain special significance if we identify it in an agent that actively deploys this covariation to representational ends. Performing such a function would, arguably, require the potential use of the internal parts for action guidance in the absence of an ongoing connection to those external parts which it they have come to resemble through a prior process of perception-action coupling (Grush, 2004; Haugeland, 1991). This detachability is not something that can be modeled in terms of a single, simple active inferrer, which is defined entirely in terms of its coupling to a target system. It is, however, potentially addressable in terms of hierarchical systems, like PP, where the activity of each ascending level can be viewed as an individual free energy minimizer, increasingly detached from current sensory input from the overall system's periphery (Corcoran et al. 2020; Pezzulo, 2017).

Still, my focus here is not on whether or not some particular free energy minimizing systems may, or may not, be able to meet the detachability criteria for having representational parts. The issue at hand is to identify what, if anything, free energy minimization as a "principle" contributes to the bioenactivist desire for a naturalistic explanation of the emergence of intentionality and agency in terms of biological autonomy. If the FEP is supposed to be something that "provides an implementation of enactivism, and in a sense supersedes or absorbs classical (i.e., autopoietic) formulations" (Ramstead et al. 2021, p. 59), then it cannot be the observer-imposed selection of a Markov blanket. Above all else, autonomy and autopoiesis are supposed systems that are *self-distinguishing* and *self-producing*. So, if anything in the FEP is able to demarcate this operationally closed and autonomous unit then, it must be the cycle picked out by the ESIA equations of the overall steady-state system.

Perhaps once this is identified and we have a naturalized account of that system's primary function, then Markov blankets might be used to formalize this hierarchical detachment and the development of proper representations. That is, they may be a part of how we build a bridge between modeling and inference with basic processes of autonomy and self-preservation. The prior question, however, is whether the steady-state ESIA cycle does a good job of describing that basic autonomy.

7 Seeking Closure

The idea that the thermodynamic openness identified by von Bertalanffy (1968) as a condition of life must be paired with some form of closure by which the individual organism may be identified and distinguished from its environmental backdrop is far from unique to either the autopoietic tradition or the FEP. Closure is central to Jean Piaget's (1971/1967) work on the nature of life; Howard Pattee (1982) proposes the notion of semantic closure; Robert Rosen (1991, 1999) draws on Aristotle to support an analysis of living systems in terms of "closure to efficient causation"; Stuart Kaufmann (1986) develops an account of "catalytic closure," later extrapolated to the more general notion of "work-task closure" (Kaufmann, 2000) and further developed by Montèvil and Mossio (2015) and Moreno and Mossio (2015) in terms of "closure of constraints."

While this enduring concern with some sort of "closure" suggests that we might be on the right track in taking it to be central to a theory of life, the variety of forms in which it has been defended should also alert us to the possibility that the closure of the FEP, and that of the bioenactivist, may not be the same thing. This is important because, as we'll now see, the steady-state ESIA cycle is far too general to capture anything distinctively autonomous about living systems, being as it is applicable to any coupled system whose collective dynamics converge to a steady state.

7.1 ESIA-Closure

We have already discussed how a Markov-blanketed set of internal states can be described as exhibiting a balance of (conditional) *informational* closure with *causal* openness—and how trivial this particular sort of closure is. The ESIA cycle introduces a new kind of closure proper to the overall steady-state system *x*, which the Markov blanket would partition into internal and external parts. The state of each component in this ESIA cycle is partly set by a deterministic function whose domain of inputs is limited to the state of another subset of components that are part of this same cycle, and partly by a Gaussian noise term *w*. This prohibits any non-symmetric biasing influences from processes that are external to this cycle, which would drive it away from its fixed steady state attractor.

As regards Varela's "intriguing paradox" and the necessity of pairing closure with an openness to environmental interchange, the ESIA cycle does admit external influence in the form of these stochastic elements of the steady-state equations. This grants a limited contribution from the environment as a source of random, symmetrically distributed perturbations away from the states prescribed in the deterministic component of the ESIA equations. Such perturbations are counteracted by the "dissipative" element of the ESIA flow, which is set so as to match the magnitude of said fluctuations (see section 6.1). The justification for this stipulation is that if this flow did not adequately counteract said fluctuations, then surprisal would not be minimized, the system would eventually drift away from its characteristic steady state, and so, according to the free energy principle, would cease to exist.

It seems worth clarifying that the kind of closure described by the ESIA cycle cannot be intended as the requirement that the particular components of some system be isolated against any *potential* influences that would not conform with these equations. There is nothing, living or otherwise, whose states are immune from the possibility of alteration by anything other than uncorrelated and symmetrically distributed fluctuations. If this were what ESIA-cycle closure required, then it could only be an idealization that is never actually satisfied in reality—which would make it a poor criterion for the presence of life.

Instead, this ESIA-closure must be understood as the definitional criteria for identifying a particular process as taking place, rather than a constraint on the range of possible interactions that the realizer of said process may be subject to. So, to say that, for instance, two coupled pendulums can be modeled in terms of a steady-state ESIA cycle does not mean that it's impossible to grab one pendulum and force it into a particular position. What it means is that such a deterministic influence would be incompatible with their entrainment and would violate the closure that defined them as "*coupled pendulums*." While the process must exhibit closure for so long as it continues to operate, its realizer remains vulnerable to influences that would violate said closure and thereby terminate that process—in much the same manner as organisms, in the process of living, are vulnerable to the possibility (the eventual inevitability) of dying.[1]

So, the requirements of steady-state and the ESIA equations define a cyclical process that is closed to any outside influences that would change its average behavior, disrupt its steady-state dynamics, and thus break the stationarity of the joint probability distribution of its generative model. It is, however, open to brief fluctuations consistent with the variance of this probability distribution. Where might we find such processes?

Unfortunately, for the FEP's pretensions to provide something that "defines the form of life that an organism is seen to enact" (Ramstead et al. 2018, sup matt 4, p. 33), something that is sufficient to characterize the emergence of adaptive behavior (Friston, 2013), and that "formalizes," "redescribes," or "subsumes" the concept of autopoiesis (Allen & Friston 2018; Korbak 2021; Ramstead et al. 2021), all that is necessary to meet these requirements is mutual entrainment among a pair of coupled systems whose collective dynamics synchronize to a steady state—systems that could then be decomposed, via a Markov blanket, into external, sensory, internal, and active parts, dependent on which part is first chosen as internal.

As Kirchhoff et al. (2019) discuss, such a description applies perfectly well to Huygens's classical example of the mutual entrainment of the two oscillating pendulums, coupled via vibrations transmitted through a connecting beam that is (or may be modeled as) a Markov blanket between them. Similarly, Baltieri et al. (2020) give a more detailed treatment of how a Watt governor coupled with a steam engine–powered flywheel satisfies the "existential triad" of the FEP, allowing their redescription in active inference terms.

If we take the arm angle of the governor as an internal state, then the free energy framework allows us to class the rotational speed of the axle, on which this directly depends, as a "sensory" state. As this increases beyond some set threshold, the arms rise as a consequence, resulting in the closing of the steam engine throttle valve, the "action" node. This reduces the flow of steam to the engine, thereby slowing its speed, which constitutes the "external" state that the governor is regulating.

On this basis, an active inference interpretation allows us to describe the Watt governor as inferring the speed of the steam engine, and the angle of its arms as constituting a model of the same. This interpretation, as Baltieri et al. note, depends on the arbitrary selection of the governor's arms as our internal state. Nothing in this setup, or the toolkit of the FEP, prevents us from selecting the engine's speed as our internal state instead and then analyzing this as inferring and modeling the Watt governor.

This triviality with which the FEP's requirements are satisfied by nonliving systems is not a unique problem for it as a theory of cognition. The idea that Watt governors and other stabilizing systems might share important principles with action-perception coordination is, after all, the basis of the dynamical approach to cognitive science (Van Gelder, 1995). To avoid attributing cognitive capacities to simple mechanical regulators, FEP-theorists could allow that a system's being describable in terms of inference does not amount to its actually performing inferential operations, as Wiese and Friston (2021) suggest. This would be to concede that the FEP does not constitute "mark of the cognitive" or an answer to "What is cognition?" any more than the idea of a dynamical system does. Rather a more modest alternative would be to suggest that, like dynamical systems theory, the FEP provides a set of tools for analyzing things in general that *may* allow us to identify the specific properties that distinguish cognitive ones (Andrews, 2021).

If, however, we were to stick with presenting the FEP as an attempt to specify the conditions under which "life—or biological self-organization—is an inevitable and emergent property" (Friston, 2013, p. 1), then this broad applicability to systems that are obviously not alive, and show no tendencies of becoming so, looks fatal.

Still, fatal for whom? If the free energy framework has indeed delivered a satisfactory formalization of the bioenactivist's autonomy-based definition of life, then it would have shown that autonomy cannot do the work that bioenactivists need it to do. The inadequacies of prior characterizations of both autopoiesis and autonomy to pick out the essential features of living systems have been extensively discussed (Bickhard, 2000; Di Paolo, 2005; Bitbol & Luisi, 2004; Collier, 2004, 2008; Bourgine & Stewart, 2004; Fleischaker, 1988) (see Froese & Stewart [2010, p. 9] for a comprehensive overview). The FEP's purported formalization cannot be blamed for its triviality if this merely reflects the incoherence, as Ashby claimed, of truly self-driven behavior of the sort the bioenactivist seeks. Perhaps all this reflects is that there is no such thing as an immanent teleology of organisms, no fundamental distinction between a system that displays "life*like*" behavior and that "*appear[s]* to actively maintain its structural and dynamical integrity" versus a truly living system that genuinely works toward its ongoing existence.

Alternatively, the fault may lie with the FEP's existential prerequisites of steady states and ESIA cycles. Even if they were adequate to formalize early definitions of autonomy, which I shall shortly argue they are not, bioenactivism is not wedded to these particular formulations. In recent years, much exciting work has been done on addressing inadequacies in these earlier accounts, to develop and refine the concept of autonomy. This work has been roundly ignored in the free energy literature. I will return to this in chapter 10 to look at the state of the art when it comes to biological autonomy. First, I will take a closer look at the history of bioenactivist formulations of closure and how these relate to the FEP.

7.2 Bioenactivism, Autonomy, and Closure

In the introduction to bioenactivism in chapter 1, I gave a brief summary of the role closure plays in Di Paolo and Thompson's (2014) definition of autonomy, where closure is realized by a network of processes such that each process is both enabled by, and a condition for, at least one other process in that network and would run down absent this network's support. Once we have selected our target of investigation, identifying these mutual dependence relations allows us to determine the boundaries of this network, in terms of those connected processes that meet these criteria.

This description of closure presents two immediate contrasts, in terms of both relations and relata, from the closure of the ESIA cycle. Firstly, where the ESIA cycle described a single process in terms of a closed cycle between changes in *states of variables*, in Di Paolo and Thompson's (2014) definition, closure is specifically a relationship *between processes*. Secondly, for Di Paolo and Thompson, it is key that this "dependence" relationship is not merely one of causal or statistical influence, in the sense that the state of a barometer depends on the air pressure, but existential dependence where one process would cease to exist as a process altogether if not for its enablement by further processes making up said network. With the further requirement that there are no additional redundancies that would compensate to continue this process in the absence of its support from within the network, then we have a system that is precarious and so, for Thompson and Di Paolo, autonomous.

The ESIA cycle, as a cycle of patterns of influence between states rather than of dependence between processes, seems to get us no closer to this formulation of autonomy than the steady-state Markov blanket did. Still, over the history of attempts to extract the key principles of autopoiesis and extend them to autonomy, one can find a wide variety of subtly different formulations of the particular kind of closure that enactivists have taken to be the foundation of autonomy.

More importantly, as Bich and Arnellos (2012) argue, there are (at least!) two distinct notions of closure that play importantly different roles in both Maturana and Varela's work on autonomy and autopoiesis—namely operational versus organizational closure.[2]

These are standardly run together not only in the free energy literature but also in bioenactivist discussion as either a mere terminological choice (Barandiaran, 2017) or as emphasizing different aspects of the same sort of system (Thompson, 2007). Somewhat confusingly, the term that is generally favored in attempts to characterize autonomy is "operational closure," with precariousness stuck on as an additional requirement. In spite of this terminological choice, the kind of requirement that contemporary bioenactivists (e.g., Thompson and Di Paolo (2014) actually describe corresponds better to the kind of generative dependence between precarious transformation processes that Varela describes by the more specific notion of *organizational* closure:

> We shall say that autonomous systems are organizationally closed. That is, their organization is characterized by processes such that (1) the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and (2) they constitute the system as a unity recognizable in the space (domain) in which the processes exist. (Varela 1979, p. 55)

It is this kind of closure, which Bich and Arnellos suggest, captures the essential element of self-production identified in autopoiesis. This appreciation for the centrality of self-production, in terms of the mutual dependence of precarious elements, is not only a feature of autopoietically informed accounts of autonomy but also of associated theories of what Letelier et al. (2011) collectively term "metabolic closure." It is the primacy accorded to metabolism that sharply distinguishes all of these above from the homeostasis-focused approaches of Friston and Ashby, whose accounts are unable to capture specific requirements introduced with the kind of preservation that is dependent on both continuous material turnover and phase transitions between different steady states—as I will discuss in sections 8.2 and 8.3, respectively.

I plan to argue that a formulation of biological autonomy that is not subject to the kind of trivializability that we have met with in the discussion of the FEP depends on this self-production that is not captured by operational closure. It is self-*production* that is not only unformalized but unformalizable in the free energy framework, and which distinguishes between a system that is merely preserved and an agent that preserves itself. In chapter 10, I will discuss more recent attempts to formalize it beyond the molecular level of autopoiesis. First, however, I will now turn to the notion of *operational* closure and the interpretation of autopoiesis as a form of homeostasis to describe both how the FEP describes this and why bioenactivists have largely abandoned such an approach.

7.3 Operational Closure

Operational closure is defined by Bourgine and Varela (1992) as follows: "A domain K has closure if all operations defined in it remain within the same domain. The operation of a system has therefore closure, if the results of its action remain within the system" (Bourgine & Varela, 1992, p. xii).

As stated, the inspiration for this definition is closure in the algebraic sense, where the range of possible outputs of some operation is a subset of its domain of inputs. While Bourgine and Varela's transposition of this to the action of a system in the second part of the above quote sounds like the requirement that this system has no effects outside itself, this cannot be what is intended, particularly given their stress on avoiding the suggestion that any system is isolated from interaction with its external environment. Rather, the point is that, insofar as these side effects do not influence the state of the system itself, they are of relevance only to us as external observers. They have no bearing on the constitution of the system and its closure of operation, which must always bring the system back to the domain in which the operation began.

Maturana (1970) described this same property more explicitly as follows:

> Living systems as units of interactions specified by their condition of being living systems cannot enter into interactions that are not specified by their organization. *The circularity of their organization continuously brings them back to the same internal state (same with respect to the cyclic process).* Each internal state requires that certain conditions (interactions with the environment be satisfied in order to proceed to the next state. Thus, the circular organization implies the prediction that an interaction that took place once will take place again. If this does not happen the system disintegrates; if the predicted interaction does take place, the system maintains its integrity (identity with respect to the observer) and enters into a new prediction. (p. 3) [my emphasis]

So, operational closure requires, firstly, that the state of the system is at each point (at least partially) determined by its own operations and, secondly, that as the range of potential states of the system is limited by its organization, so is the range of interactions into which it can enter. In Maturana and Varela's writing, possible states are referred to (somewhat confusingly) as potential "structures." Whether these structures are possibilities, or not, is dependent on whether they preserve the higher-order relational organization that defines the system as the particular operationally closed system that it is, even as its constituent parts may change. Just as the level of a mercury thermometer may rise and fall while it retains the organization that defines it as a thermometer, so a cell may lose and gain specific molecules that constitute it, while still retaining the same overall relations between its molecular network. As such, when treating autopoiesis as defined by operational closure, Maturana and Varela declare that "it is thus clear that the fact that autopoietic systems are homeostatic systems which have their own organization as the variable that they maintain constant" (Maturana & Varela, 1980, p. 80).

Like the pronouncements of the FEP, operational closure has the feeling of triviality. Such an abstract formal specification grants no insight into the methods by which ongoing existence is achieved, it makes no reference to the particular thermodynamic considerations peculiar to living systems, and it does not distinguish their capacity for the turnover of their material components from the changes in state of a common mechanism. The operations of all systems will also have some effects that feed back to the system itself, and all systems are limited in the range of states they can occupy. As an analysis of a living organism then, operational closure alone feels explanatorily impoverished.

I have already mentioned that operational closure is not intended to suggest that the operations of the system do not have any effect on the world outside said system. Rather, the point is that such effects belong to the domain of external observers and are not a part of the constitution of the system itself (Maturana, 1975; Maturana & Varela, 1980/1970; Varela, 1979). As Maturana (1975) describes:

> Given a closed system, inside and outside exist only for the observer who beholds it, not for the system. The sensory and the effector surfaces that an observer can describe in an actual organism, do not make the nervous system an open neuronal network because the environment (where the observer stands) acts only as an intervening clement through which the effector and sensory neurons interact completing the closure of the system. (Maturana, 1975, p. 318)

Just as outputs are off-limits, so too, as Thompson (2007) explains, is talk of inputs, at least "in the usual sense." This caveat is particularly important. While outputs that do not loop back to affect the system are, by definition, irrelevant to a characterization of its intrinsic properties and detectable only by an additional observer, no system is isolated from any environmental influences on its state. Such "inputs" do not require an external observer to detect, so what justifies their exclusion?

The answer is the concept of structural determination (Maturana, 1975) or self-determination (Varela et al., 1991), a corollary of operational closure that describes how, as far as the system is concerned, any environmental influence will appear only as "perturbations within the processes that define its closure, and thus no 'instructions' or 'programming' can possibly exist" (Varela 1979, p. 58). What the notions of operational closure and structural determination pick out is not an isolated system, closed off from any interaction with its environment, but rather a consequence of the particular descriptive perspective taken when we describe the internal constitution of the system in terms of its operationally closed organization. When the system is described in such terms, an environmental change does not "instruct" the system by entailing a particular output but perturbs it, with the consequences being dependent on the state of the system at the time of the input and how the system's closed operations work to preserve its fixed organization.

As Maturana (1970) describes, any event that does not merely perturb those operations but causes them to cease altogether (as in the case of fixing one coupled pendulum to a set position) would correspond to the breakdown of its organization and the destruction of the system. The FEP's steady-state ESIA cycle can be interpreted as a transposing of the concepts of operational closure and a fixed organization that may vary in state, into the language of random dynamical systems and statistical models of the same. Like the idea of "homeostasis of organization," the FEP abstracts away from turnover of components, treating these as equivalent to changes in state that are compatible with the preservation of the operationally closed organization, specified by the ESIA cycle. In arguing that the FEP can serve as a formalization of autopoiesis, Wiese and Friston (2021) thus draw specifically on this characterization of the latter in terms of operational closure and homeostasis of organization.

In the stochastic differential equations that the FEP took to define the overall steady-state system (equations 4 and 5, in section 6.1), we find a similar move to demote environmental influence to perturbation. This is the distinction between the ESIA-internal variables of the deterministic component that make up x, define our "operationally closed" system and determine its steady-state dynamics, versus the noise terms (w) that are supposed to encapsulate all outside influence.

Like the FEP's steady-state ESIA cycle, this characterization of autopoiesis and autonomy in terms of operational closure and homeostasis of organization is extremely minimal as a characterization of a living system. As I will argue in section 7.5, it is no longer popular as an analysis of autonomy and is insufficient to serve the aims of bioenactivism.

Still, it is in the formulations of operational closure and structural determinism from Maturana and Varela's writings in the 1970s and 1980s, and in the FEP's steady-state ESIA cycle, where I think the paths of Friston et al. and the development of the bioenactivist viewpoint come closest to crossing. As I will argue in the next section, however, they quickly diverge.

7.4 From Freedom and Stability to Dependence and Purpose

As with the FEP's steady-state ESIA cycle, the properties of operational closure and structural determinism are not unique to living systems. But then neither did Maturana or Varela intend to take them as such. The autopoiesis of a single cell is an instance of operational closure, but that does not mean to say that this kind of closure exhausts all that is of interest in the autopoietic characterization of life. As Villalobos and Ward (2015) point out, the examples Maturana (1987, p. 73) cites in illustration of structural determination are not biological but mechanical or computational cases like a washing machine or a lightbulb. Similarly, while operational closure may be easiest illustrated with classical dynamical systems examples, such as the Watt governer or the thermostat, computational systems can also be analyzed in terms of a closed loop of interaction with their environment. A Turing machine, for instance, writes to and reads from the same tape, allowing its "output" at one time to alter its "input" at the next (Villalobos & Dewhurst, 2018).

So, what of our hopes to naturalize teleological or intentional talk in the autonomy of living systems? If operational closure as a description of the homeostasis of organization were indeed all there is to autonomy, and this is satisfied by anything with a feedback loop through its environment, then we would have to attribute goal-directed behavior to toasters and to teamakers. This is not, however, a view that should be attributed to either Maturana or Varela. The development of such an implausible position can only come from ignoring how their views developed and diverged and, crucially, the mistake of conflating together the relatively trivial properties of self-determination and operational closure with the more demanding requirements of self-*production* and the closure of precarious processes that are at play in more recent bioenactivist attempts to naturalize teleology.

In Maturana and Varela's earlier work, their shared stance is explicitly anti-teleological. As Maturana describes the guiding tenets for his work in the (sole-authored) introduction to *Autopoiesis and Cognition: The Realization of the Living*, "notions of purpose, goal, use or function, had to be rejected" (Maturana & Varela 1972/1980, p. xiii). Chapter II, "Dispensability of Teleonomy," is dedicated explicitly to this purpose, wherein they describe a perspective on the organism completely at odds with bioenactivism's phenomenologically motivated concern for sense-making, immanent teleology, and intentionality:

> Since the relations implied in the notion of function are not constitutive of the organization of an autopoietic system, they cannot be used to explain its operation. The organization of a machine, be it autopoietic or allopoietic, only states relations between components and rules for their interactions and transformations, in a manner that specifies the conditions of emergence of the different states of the machine which, then, arise as a necessary outcome whenever such conditions occur. (p. 86)

As with Ashby then, the concern is to exorcize an internal goal-directed driver of behavior, a primitive volition standing outside of ordinary causal entailment, from our explanation of living systems. If one is content with this view that there is nothing especially teleological or purposeful about an organism and believes that *if* there is even such a thing as "original intentionality" it does not begin with the first autopoietic cell, then operational closure, homeostasis of organization, and indeed free energy minimization might appear to serve as an adequate characterization of living and the nonliving systems alike.

Varela himself appeared unable to remain content with this anti-teleological position for long. In *Principles of Biological Autonomy* (1981), he already begins to negotiate room for teleological talk as part of a pluralistic view of explanations, treating it as an alternative form of description that an external observer might validly use to better understand a system's behavior—in much the same sense as talk of "symbols," "inputs" and "outputs," or "the environment" may be proscribed from an operational characterization of a system's intrinsic properties, while being perfectly legitimate at the level of communication about that system between external observers.

This is not yet the bioenactivism that I am looking for. The view that teleological explanations and intentional attributions may be legitimate in certain contexts is just as compatible with an instrumentalist characterization that takes the difference between organisms and mechanisms to be one of complexity, and the appropriateness of such terms to be a matter of their success in abstracting from this detail to yield the kind of predictive regularities that interest us.

By 1999, however, Varela was influenced by Kant and Jonas's work on the idea of organisms as "natural purposes" and began exploring the concepts of "original intentionality" and "sense-making" as unique to life, coming around to the position that these do lead to the reintroduction of a kind of teleology that is "intrinsic to life in action" (quoted from an email exchange in Thompson, 2007, p. 453–454). This about-turn culminates in a 2002 article with Andreas Weber that marks the definitive break with Maturana's (and Ashby's) attempt to treat living and nonliving systems as purposeless entities alike.

As Weber and Varela (2002) argue, drawing extensively on Jonas (1953), the distinguishing aspects of the living can be stated as follows:

1. It exchanges its matter and acts thereby from a subject pole partially independent of the underlying matter.
2. As precarious existence, it is always menaced by concern (Sorge), the need to avoid perishing, and to do this, it is again completely dependent on matter whose characteristics are the reason for its concern.
3. Already the simplest forms of life have thus a subjective perspective as a result of this existential need. Therefore:
4. Life as such will always be captured in the antinomies of "freedom and necessity, autonomy and dependence, I and world, relatedness and isolation, creation and mortality" (Jonas 1973, p. 3—quoted in Varela & Weber, 2002).

These are not consequences of operational closure, homeostasis, and structural determinism, properties that autopoietic systems share with systems in general, but rather they express the ways in which an organism's self-producing character creates a difference in kind between the living and nonliving. For Jonas, the key to all of these properties is an understanding of precarious self-production through metabolism—the dependence of living systems on continual flows of matter in order to preserve and rebuild their precarious organization. Unlike in homeostasis, the key feature of metabolism is not just preservation *in spite of* material turnover, where this is presented as perturbation within a homeostatically conserved organization, but precarious *dependence on* that turnover. Or, as Jonas (2001/1966) nicely puts it, the distinguishing feature of the organism is its relationship of "needful freedom" with matter.

For Weber and Varela, it is this needful freedom of metabolism, as opposed to the mere freedom of homeostasis-of-organization, that is now presented as the key feature of autopoiesis, rather than an accidental aspect that should be abstracted away from in the more general notion of autonomy. Homeostatic systems are fundamentally passive ones that act only in response to external disturbance and are capable of remaining stable not only *in spite* of such disturbances but also in the total absence of them. In contrast, a metabolic self-producer is necessarily active and *dependent* on externally introduced inputs to replenish and refuel its continuation. It is this that makes a network of chemical reactions a self-producer, and it is this precarious dependence that is key to the bioenactive naturalization of goals, intentionality, and immanent teleology.

The key point here is that, as restrictions on how a system can change *if* a change state occurs, neither homeostasis, operational closure, nor the steady-state ESIA cycle describes something that is necessarily *dependent* on these changes for its existence. All that is necessary for a system to be an ESIA cycle is the conditional requirement that *if* one component changes state, then it must be the result of a change in another component of the appropriate set. So, *if* an internal variable changes, then there must have also been a change in either another internal, sensory, or active variable that caused it—and these changes must conform to the dynamics of our steady-state equation so as to preserve a stationary probability distribution. But, crucially, there is no absolute requirement for ongoing change to preserve either the individual components or the overall ESIA-cycle organization.

Take the coupled pendulums when they are at rest. Insofar as their average state remains fixed at equilibrium, they will satisfy a steady-state requirement for a stationary probability distribution, used as the generative model of active inference. At rest, they have minimized their surprisal quite perfectly and, should they happen to be perturbed by external noise, they can be expected to minimize the resulting unlikeliness of being out-of-equilibrium with perfect alacrity. Insofar as the position of one pendulum cannot directly change the velocity of the another without changing the state of the beam, their potential patterns of influence will also conform to that of an ESIA cycle and the beam will qualify as a "Markov-blanket" between them. Still, the beam itself, like all other parts of the system, is perfectly self-sufficient. It will neither crumble nor dissipate if pendulums cease to oscillate.

So, the kind of system described by a steady-state ESIA cycle is one where we have some fixed constraints that determine the dynamics of the system, but where these constraints do not depend on those dynamics in turn. The fact that the beam does not depend on the movement of the pendulums can seem hard to miss, and yet we find supporters of the free energy framework claiming that "the Markov blankets are a result of the system's dynamics. In a sense, we are letting the biological systems carve out their own boundaries in applying this formalism. Hence, we are endorsing a dynamic and self-organising ontology of systemic boundaries" (Ramstead et al., 2019. p. 3).

This is presented as describing the same existential dependence between dynamics and structure as found in the cell membrane, but as we've seen with the coupled pendulums, this is not a general feature of Markov blankets or ESIA cycles. The cellular membrane is an intrinsically unstable configuration, and if the internal workings of the cell do not act to replenish this boundary, it will dissolve. The same is not true for the blankets in Huygens's pendulums, nor for the coupling between the Watt governor and the steam engine.

A probabilistic graph and its attendant Markov blanket merely describe how a system's structure constrains its dynamics; they do not mandate any reciprocal dependence of this structure on those dynamics. At best, the relationship we might have here is an epistemological one, in as much as the dynamics reveal the independencies of a boundary. Still, given that conditional independence is everywhere, and those particular boundaries the FEP chooses to focus on are typically things that we have picked out by means prior to any measurements of statistical relationships, it does not seem a particularly informative clue to the selection of boundaries.

7.5 Self-Production Is Not "Homeostasis of Organization"

There are, as Di Paolo et al. (2022) note in their enactivist critique of the FEP, other reasons to reject the deflated account of autopoiesis-as-homeostasis. For one thing, as I will describe in chapters 8 and 9, a distinguishing feature of organisms is the difficulty, and I will suggest the impossibility, of identifying any invariant organizational feature that is both (a) specific enough to individuate a particular organism from others and (b) something that *must* be preserved throughout its all of its open-ended developmental possibilities. Another, more basic, problem with this interpretation, as Di Paolo (2005, 2010) describes, is that homeostasis allows for *variation* around a stable point, whereas autopoiesis is a binary property that does not admit of graduation. One is either a self-producing network or one is not, and once this breaks down, it's too late to do anything about it.

For the purposes of explaining why the FEP cannot serve the aims of a bioenactivist, the main point is nicely expressed by Mossio and Bich (2017) with regard to the failings of homeostasis-based accounts of the organism more generally:

> It presupposes the existence of the organisation that under certain circumstances it contributes to maintain stable. In particular, homeostasis does not capture the most distinctive generative dimension of biological organisation, i.e. the fact that the components involved in feedback loops are not only stabilised, but produced and maintained by the very organisation to which they belong. In a word, homeostasis misses precisely self-determination. (p. 1096)

And, as they continue:

> Technically, the "goal" of a homeostatic mechanism is defined as the interval within which the mechanism maintains the target variables. Yet, it does not make any difference from the point of view of the definition whether the interval is extrinsically established by a designer, as in the case of artefacts, or intrinsically identified with the conditions of existence of the system, as in the case of biological systems. Both cases can pertinently be said to be homeostatic. However in failing to account for their difference, Cybernetics misses the crucial dimension of biological teleology. (p. 1096)

A homeostatic description can only come after we have determined some set of variables and the state at which they must be preserved. As such, it is indifferent to the means by which these essential variables and their bounds of viability are determined—whether by an external designer or by something intrinsic to the homeostatic system itself.

The adoption of a Jonasian approach to living systems in Weber and Varela (2002), subsequently carried through in Thompson (2007) and Di Paolo et al. (2017), represents a decisive break with the attempt to treat autonomy as a generic property of operationally closed systems, to reduce autopoiesis to homeostasis, and to treat organisms in the same terms as any other physical system. Awareness of this diametric opposition between Maturana (and Ashby) on the one hand, and bioenactivism on the other, is essential to a coherent understanding of either (Villalobos, 2013; Villalobos & Ward, 2015). It is not that the former neglect this intrinsic teleology, such that their work can be straightforwardly supplemented by an account of it. In defining autonomy in terms of the generic properties of homeostasis of organization and operational closure, Maturana and Ashby debar such a possibility. Lumping these views together, as the free energy literature has tended to do, creates an incoherent frankentheory that no amount of mathematical stitching can hold together.

If the FEP's steady-state ESIA cycle is only a description of a specific form of operational closure, then it can hardly serve the bioenactivist as an analysis of the features in virtue of which living systems are (1) distinguishable from the nonliving and (2) sources of proto-intentionality. But I don't want to just reject the FEP's analysis of the organism for failing to live up to my prior commitments to the bioenactivist viewpoint. A FEP-theorist might respond that Maturana and Ashby had the right of it, that the attempt to treat organisms as unique and goal-directed is misguided and that they are just one physical system among many, where a vastly greater degree of complexity combined with our inability to fully comprehend this leads us to mistakenly attribute a difference in kind. Or they might suggest that the steady-state ESIA cycle be taken only as a partial, necessary but not sufficient, set of conditions for a definition of biological self-organization. Alternatively, they could propose that the FEP itself should not be understood as a theory about organisms at all, but rather a set of mathematical tools, a statistical redescription of dynamical systems theories that might then be used to formulate a description of those features that *are* specific to organisms.

None of these responses will work. As I will argue in the next chapter, the FEP's problem is not merely that formalization of a homeostatic identity is too generic to distinguish the particular features of autonomous biological agents. The problem is that the principles of stability that it presents as necessary are exactly those principles that biological systems are uniquely prone to violate.

8 A Theory of Everything, or Just of Every "Thing"?

Now, at last, we have a clearer understanding of how the FEP defines the existence of a system, or, in Friston's (2019) terminology, every "thing" to which the free energy principle is supposed to apply, in terms of the coupled equations of the steady-state ESIA cycle. With this in hand, we can extract the two assumptions that the FEP makes about a system.

1) Tendencies: The parts of the system tend to revisit the same state, or to cycle through the same set of states, with a frequency that does not change over time.
2) Dependencies: The interactions between these different parts of the system do not change.

To connect this to Bayesian inference, the FEP then redescribes these dynamics in terms of an invariant low-entropy joint probability distribution over the state of all of the parts of the system—which gives us the "generative model." The partition between external, sensory, internal, and active variables now becomes a statistical one—the Markov blanket. The FEP then uses this partition to interpret internal variables as "inferring" external ones, in as much as the statistical properties of the internal variables converge with those of external variables, despite their being independent of one another when conditioned on the Markov blanket of sensory and active variables.

So, there are two key moves in the free energy framework: (1) a putative definition of every "thing" that the principle is supposed to apply to and (2) the attempt to use a statistical redescription of this to license the attribution of inferential properties. The problems with this latter move, which is essentially a slide from correlation to representation, have been well discussed as summarized in section 3.5. Yet, as Raja et al. (2021) note, the first move on which it depends has too often been granted a free pass, despite the highly unlikely assumptions it makes about those entities to which the free energy principle is supposed to apply.

It's this state of affairs that has led to the FEP's being either lauded as a "first principle" or decried as "unfalsifiable." *If* this proposed definition of a system holds (along with some further technical specifications, for more on which see Biehl et al., 2021 and Aguilera et al., 2022), *then* it entails that the system will be formally describable in terms of free energy minimization. In this sense, Friston is correct to claim that the free energy principle itself is not an empirical hypothesis—for all that specific hypotheses about cognitive architecture, such as PP, may be derived from it—and, hence, frequent objections raised to its lack of falsifiability miss the mark. Yet, the FEP should not be taken as merely the working out of a tautology of existence either for all that it has been presented as such. This would presuppose that the FEP's steady-state ESIA cycle was already an accepted definition of "existence," biological and otherwise. Rather than a tautology then, Hohwy (2021) proposes that the FEP is best understood as an attempt to address this definitional gap through its putative account of what it is to be a (self-organizing) system.

The appropriate test of a proposed analysis is not whether falsifiable experiments can be derived from it, but that doesn't mean it is immune from criticism and counterevidence. Instead, we look to how well it accords with both common and scientific practice. Neither the FEP nor any other piece of formal analysis is required to submit entirely to the absolute authority of this bicameral legislature—as though either chamber were even capable of producing a unilateral ruling in the first place. A philosophical analysis may worry the fabric of our linguistic habits, to expose incoherence in the everyday applications of a concept. It may likewise criticize scientific terminology for losing touch with everyday use. If, however, the proposed criteria result in extensions of a concept that are diametrically opposed to both everyday intuitions and to our best scientific understanding, then the only conclusion is that the philosopher (or neuroscientist) is talking about something else entirely.

I have already described one problem with the FEP's definition of a system: namely that in attempting to treat both animate and inanimate entities in the same terms it rubs up against the folk understanding that there is a difference in kind between the existence of a person and that of a pendulum. In treating both as nothing other than steady-state ESIA cycles, the FEP fails to account for why we tend to talk of one as an agent and pursuer of goals and the other as a mere mechanism. As such, it threatens to lead us either into deflating instrumentalism about intentional talk, such that it depends only on the greater complexity of the former's behavior, or to inflated panpsychism in which every system in a stable coupling would count as an inferring agent.

Neither option works for the bioenactivist viewpoint, but that is not a good enough reason to reject the FEP. Physicalism and mechanism are still the order of the day in most respectable scientific circles, who are liable to turn up their noses at any whiff of "vitalism." A theory that treats the apparent gap between life and non-life as an illusion, one that stems only from the complexification of ordinary mechanisms, may be more likely to appeal.

Friston and others working on the FEP have begun to develop an account of what this complexification might look like. Suggestions include the incorporation of a solenoidal component to the system's dynamics, resulting in a limit cycle attractor, rather than fluctuations around a single point (Friston, 2019b), or hierarchical extensions that allow for temporal or counterfactual depth in a system's predictions, wherein actions are selected to minimize the long-run average of free energy over an entire trajectory (called expected free energy) rather than just to minimize immediate free energy (Wiese & Friston, 2021; Friston et al., 2020).

As I will argue in this chapter, all attempts to fix its broad applicability by casting the organism as a "special case" of a steady-state ESIA cycle will fail. As I will argue, the problem with the existential imperatives by which the FEP defines a system is not their generality but their *contingency* when applied to biological forms of existence. The stationarity of a joint probability distribution, and the stability of tendencies and dependencies that it implies, may do well enough as necessary requirements for the continuing existence of an inanimate substance. Living systems, in their temporary coincidence with a particular physical body, may even happen to meet these two stability requirements over some duration. But, as necessary principles meant to define an organism's ongoing existence, they are both false.

This is not just a flaw of the FEP and its conceptualization of the organism as a homeostatic mechanism with stable behaviors and stable parts. Instead, this misconception stems from a broader ontological framework that views organisms as substances, and specifically, as machines. But the mechanical and the substantial do not exhaust all possible modes of existence. Unlike machines, organisms may change both their components and the rules that govern the behavior of these parts in unprestatable ways, leading a number of philosophers, theoretical biologists, and complexity theorists to propose that they are much better captured by a processual ontology.

The stability of parts and properties may well be a trivial property of inanimate objects, but what distinguishes living systems, I will argue, is precisely their intrinsic instability in both material constitution and in organization and behavior. This is a difference in kind between the living and nonliving, one that can be identified independent of any folk, or bio-enactivist, commitment toward attributions of biological autonomy and which suffices to show that the FEP cannot serve as the basis for a theory of living systems—bioenactivist or otherwise.

8.1 Processes and Substances

The FEP presumed two sorts of stability in our system. The first: invariance of interactions between our parts, such that even as states of variables may change, the statistical dependencies *between* these[1] remain the same. The second: invariance of the statistical tendencies of each individual part. This allows that the state of each part may change but requires that if it has been in a particular state eight out of ten times previously, then it must also spend 80 percent of its time in that state in future. This second requirement has been justified in increasingly general ways: as a formalization of biological homeostasis in particular; as a mathematical description of stability through perturbation more broadly; and, most generally of all, as a principle of all existence.

Friston (2019a) motivates this by defining a "thing" as something that is "distinguishable in a statistical sense," cashed out in terms of having states characterized by a low entropy probability distribution that remains constant over the duration of said thing's existence. The idea being, crudely put, that if a thing did not regularly revisit the same states but rather wandered off to ever new regions of possibility then how would we reidentify it? Or, as he puts it,

> nearly every system encountered in the real world is self-organising to a greater or lesser degree—suggesting that self-organisation is, in itself, unremarkable. Put another way, if systems did not self-organise they would have dissipated before we had a chance to observe them. (Friston, 2019, p. 24)

If this were the case, if our ability to observe and identify a system over time were dependent on its exhibiting this kind of stability, then we would not need recourse to the specific importance of homeostasis to an organism to make the case that they are steady-state systems. This condition would instead be derivable from this more general requirement for what it takes to exist over time.

But is this really a requirement for everything we can identify as persisting over time? It seems to me that there are plenty of things that we can reidentify in spite of their violations of any steady-state condition. Songs have their choruses, stories have their tropes, and dances their motifs, but we can also recognize a continuity throughout the screwball transitions of Bohemian Rhapsody, we understand that when Odil ends the 32 fouettes of the Black Swan Pas de Deux to return to her partner that the dance goes on, and as the timer ticks down on the bomb handcuffed to Sean Connery's wrist, we know that this is a single ongoing moment of peril, part of a broader, non-repeating sequence that constitutes the overall narrative unit that is *Goldfinger*.

Moreover, we can continue to recognize this continuity even as the realizing parts change: as a melody moves between members of an orchestra, a relay race moves between runners, or as a river churns through water molecules. Friston (2019, 2013) takes such material churn to be incompatible with our need to identify the stable parts whose stable interactions give us a causal graph as a real entity—though as we will see in the next section, it is not clear that he is correct to do so. Given that inanimate objects may also persist through an exchange of parts, then if the FEP cannot deal with this, it would presumably raise an issue for Friston's aspirations to a theory of the nonliving also.

With regard to the violations of steady state, it would be reasonable to respond that Goldfinger, Bohemian Rhapsody, and Swan Lake are not really objects but *processes*. That repetition may be only one form of continuity in a process, but the FEP is a theory of "*things*" only in the narrower sense of a persisting substance, or more specifically, of a mechanism. As Wiese and Friston (2021) put it, these "things" are: "Systems that exist over some appreciable timespan, in the sense of having an attracting set, i.e., revisiting the neighbourhood of characteristic states despite external perturbations—of which living systems are a subset" (p. 7).

Still, if this means that the FEP is not truly a theory of every possible form of existence, then we have no *a priori* reason to accept that it characterizes the kind of existence that organisms have. To motivate this, we would need to first argue that organisms are indeed substantial "things" rather than processes, and secondly that the continued existence of a substance, or at least of a biological substance in particular, depends on the particular forms of stability that the FEP describes.

It might seem obvious that a cat is more like the ballet shoe than the ballet. Unlike Swan Lake, you can pick a cat up and throw it around—though it's inadvisable to do so. Yet, while something like mass and extension (as necessary prerequisites of pick-up-and-throwability) may be part of the ordinary conception of "substance," throughout the history of metaphysics, this term has been used in a variety of distinct technical senses (Morgan, 2021). This lack of one agreed-upon definition can make it somewhat difficult to assess the prospects of a "substance ontology" for organisms in general. For our purposes in evaluating the FEP, however, the most apposite characterization of a "substance" is what Morgan terms the "essentialist" account, in which something is defined by a particular set of properties that persist throughout its existence, allowing it to undergo only those changes that do not violate these properties.

This essentialism is the notion of substance at issue in recent criticisms from defenders of processual accounts of organisms (Nicholson, 2018; Dupré & Nicholson, 2018; Meincke 2019). As Dupré and Nicholson (2018) state, the key claim they reject is that organisms should be understood as things which persist "by virtue of their continued possession of certain essential properties, which make those things what they are and which remain unchanged over time" (p. 24). In contrast, the processual account looks not for the essential features of X, such that they could identify it in any randomly selected time-slice, but asks instead "how should I follow X through time" (Guay & Pradeu, 2016, p. 318). Rather than by an atemporal set of features, a processual identity is individuated by what Lewin (1922) terms relations of "genidetity" between each successive time slice, where the latter is a generative product of the former. Such a relationship might, for instance, be described in terms of autopoiesis, where what makes one organism-slice a continuation of the previous one is that its chemical components are the product of the synthesizing activity of the prior set.

The processualist may allow that said process can be contingently "stabilized" in a particular material "cat-substance" that we can stroke or hold but must take that particular substance as a temporary episode in the more fundamental process necessary for being a cat. Moreover, rejecting the idea of organisms as essentialist substances is not to prohibit an organism from having *any* unchanging properties, but rather it is to claim that those invariant properties are not what individuate it as the particular organism that it is. If something is a persisting substance then, when viewing it at two temporally disconnected instances, it should still be possible to reidentify it as the same thing without knowing anything about what went on in the intervening period. If it is not, then the only way to answer this question would be to follow the entire temporal trajectory from the first moment to assess whether it connects up to the second in some, as yet unspecified, "right way."

Just as there can be different accounts of the "right way" for a processualist to follow an organism through time, so the substantialist's requirement of an essential invariant still leaves open a wide range of different substance ontologies that might be offered for the organism. The free energy framework, as I have described, seems committed to two essential invariants: (1) the stability of the parts that are taken to literally instantiate the real causal graph of the ESIA cycle and its attendant "real" Markov blanket and (2) the stability of their behavior that gives us the fixed steady-state function and associated, invariant generative model.

Friston (2013, 2019), as mentioned, often talks about the first of these, the stability of parts, in terms of the stability of material components. The stability of some fixed aggregation of material stuff is perhaps closest to the ordinary concept of substance, as when we talk of a substance as something with pick-up-and-throwability or when we report that "there's a mysterious sticky substance all over the baby's highchair." An inability to account for changes in this material realization would certainly raise a problem for the FEP in light of the continual recycling of parts that occurs in all organisms. Indeed, given that we can change a graphics card, set of wheels, or planks without creating a new computer, car, or ship each time we do so, this would raise problems for the FEP's ability to serve as a theory of anything. As we'll see in the next section, this requirement of stable material parts is a strange thing for Friston to have committed himself to. The FEP might easily abandon this to talk, as most machine-style models do, in terms of the stability of formal parts.

This comes with two problems, however. The first is that in abstracting away from material turnover to focus on the stability of formal parts, we erase the difference between a structure that is stable *in spite of* possible material exchanges and a structure whose stability is entirely *dependent* on ongoing material turnover. This distinction, as already mentioned, is crucial to the bioenactive, or Jonasian, conception of the primitive intentionality of organisms.

The second problem is that this strategy relies on a different type of essential invariant: namely an invariant organization, describable by some mathematical equation that remains fixed even as the states of its variables change. It is this that allows us to individuate a formal part in terms of the role it plays in this equation, even as its material realization changes. As we will see in the second half of this chapter, however, organisms are unique in their ability to persist not only through material turnover but also through radical changes in organization. And, as I will argue in chapter 9, there is likely no level of abstraction at which we can identify an invariant equation that is both specific enough to individuate this particular organism and flexible enough to allow us to derive every change it might possibly undergo in the course of its life span.

8.2 The Instability of Organic Parts

The fact that organisms are continuously replacing their components by interchanging matter with their environment appears to pose a problem for the requirement of fixed parts, and more generally to mark at least one quite fundamental difference, both from machines and from other substances in general. To find nonliving systems with the same property, we have to look not to the mechanical but to phenomena such as tornadoes, whirlpools, rivers, and candle flames—in other words, the archetypal cases for a process ontological perspective.

Friston (2013, 2019b) explicitly cites such material turnover as incompatible with how the FEP defines a system, but then quite bizarrely presents this in support of the framework's particular suitability for living organisms. A principal reason for taking a single-cell and its membrane as the canonical Markov-blanketed free energy minimizer, he claims, is the very stability of its components, contrasted to a candle flame, which

> cannot possess a Markov blanket, because any pattern of molecular interactions is destroyed almost instantaneously by the flux of gas molecules from its surface. Meaning we cannot identify a consistent set of blanket states rendering some internal states independent from other states (Friston, 2013, p. 2)

And as he repeats, in a more recent extended monograph on the free energy principle,

> it does not easily accommodate the fact that the particles that constitute a Markov blanket can, over time, wander away or, indeed, be exchanged or renewed. The canonical example here would be the blanket states of a candle flame, whose constituent particles (i.e., molecules of gas) are in constant flux. (Friston, 2019a, p. 50)

To cite the membrane as a point of contrast is an odd choice, given that its constituent parts are continuously consumed and regenerated by the cell's metabolic network. Membrane turnover via endo- and exocytosis is a means for all sorts of self-organizing behavior, from regeneration and growth to the transportation of molecules between the interior of the cell and its environment. In the cellular slime mold *Dictyostelium*, for instance, membrane turnover has been proposed as a mechanism of locomotion, with estimated times for complete turnover in the order of four to ten minutes (Aguado-Velasco & Bretscher, 1999). Within the cell too, amid the "internal states" presumably realized by particular proteins, we find turnover times much shorter than the life span of the overall system—on the order of around two days for a non-dividing mammalian cell (Toyama & Hetzer, 2013). This flux continues up to the multicellular level where, in the human body there is an estimated daily turnover of around 0.2 percent of total cellular mass (Sender & Milo, 2021). As the physiologist John Scott Haldane, one of the earliest "processualists" of twentieth-century biology, describes, "the organs and tissues which regulate the internal environment . . . are constantly taking up and giving off material of many sorts, and their 'structure' is nothing but the appearance taken by this flow of material through them" (Haldane, 1917, p. 90).

Such a porous and protean thing is much more like a candle flame than it is like the ESIA cycle's fixed patterns of interactions between fixed parts.[2] If the basic units of our causal graphs are the states of token particles, as Friston (2013) takes them to be, then their statistical dependencies will lack the stability necessary to establish patterns of conditional independence between them, and for the identification of a Markov blanket between unchanging sets of "internal" and "external" components. In the cell, a previously "external" molecule is free to waltz right through its membrane to start interacting directly with an "internal" one, blithely violating the FEP's basic definition of a system.

This is especially problematic for a realist about ESIA cycles and their Markov blankets, who treats them as something that the organism literally instantiates and that makes it what it is. Such a realist will find that the components of any "real" Markov blanket they identify around an organism will dissipate on timescales that are shorter than that life span of the organism whose "very existence," they claim, "depends" on that boundary's preservation (Allen & Friston, 2018).

Still, one might think that what matters is not the stability of interactions between component particles, as Friston (2013, 2019a) seems to take it, but rather the stability of the higher-level organization. As noted in chapter 5, the realist about probabilistic graphs need not hold that they are instantiated at the level of interactions between the states of particular token particles. We might instead take features of the cell's macroscopic organization, such as intracellular and extracellular glucose concentration, and describe how one cannot affect the other without a change in the state of transmembrane channels. The movement of a particle would thus correspond to a change in the state of some more macroscopic fixed node, rather than the breaking and creating of new connections in a particle-level causal graph.

The idea that we can abstract away from material turnover is, after all, the key idea behind the notion of homeostasis of organization that Wiese and Friston (2021) attempt to use as a bridge between autopoiesis and the free energy theory. The point of defining the fixity of the organism, as Maturana and Varela (1980/1972) did, in terms of organization, rather than the variable material instantiations of this, is that it allows us to talk of an organism like a machine, in terms of its fixed formal parts and the rules governing their behavior.

This machine-substance view of what it means to have fixed parts not only accommodates material turnover in the organism but arguably better captures how this requirement is supposed to apply to machines too. Machines may not exchange their material components as a matter of course, but they do admit of such exchanges. Unscrupulous salesmen aside, when we repair and replace the wheels on a car, we are not inclined to say we have created a new machine. While having wheels is a fixed invariant of a car, having some particular set of wheels is not. From the perspective of the organization level, the old wheels and new wheels qualify as the same formal part. As described in the previous section, the fixity of material realization is only one possible view of what the essential invariant of a substance is. In both the organism and the machine case, however, the substantial invariant is better captured not by fixity of material properties and parts, but by fixity of formal ones.

So, we cannot take our parts for granted as something like specific atoms or molecules. To construct a fixed graph that abstracts away from this constant turnover of microphysical entities, we need to describe the invariant form of our particular living system such that we can identify its fixed formal parts. These parts must be individuated prior to being able to create a graph of the connections between them. As such, neither the ESIA cycle nor a Markov blanket is of any use in individuating them, for in order to construct these we need to already have divided our system into fixed units, such that we can then assess any relations of dependence or independence between them.

There are bigger problems for the FEP here than its inability to serve as first principles analysis. Bigger problems, that is, than its requirement for some prior specification of the fixed organization of the organism, from which the parts needed for its analysis in terms of generative models and causal graphs might subsequently be derived. As I will discuss in the second half of this chapter, organisms, unlike machines, persist not only through material turnover but also through radical transformations of organization. Such transformations, I will suggest, cannot be described in advance by *any* invariant set of equations that are both sufficiently flexible to derive these transformations and specific enough to individuate that particular organism. Without this, it seems unlikely that we can pin the organism to an invariant set of formal parts any more easily than we could tie it down to material ones.

Putting this aside, temporarily, let's presume that we *could* give an account of how to identify fixed formal parts for any organism amid its material turnover, in order to treat it like a machine and so to redescribe it in the FEP's terms. Even if we could, as I will argue in the next section, in doing so, we miss out on a fundamental distinction between the organic and the mechanical in terms of how the former alone *depends* on this material turnover for whatever temporary stability its formal parts may have.

8.2.1 Why Metabolism Matters

To say that organisms, like machines, are substances in the sense of having an invariant organization and formal parts allows us to abstract away from material turnover, without outright denying its occurrence as Friston (2013, 2019) appears to do. But should we abstract away from this turnover? We've already seen that the bioenactivist, following Jonas (1953), would be inclined to protest such an abstraction as erasing the distinction between a formally defined machine that can admit of such exchanges, versus a precarious organism whose existence *depends* on them. Even those who reject the idea that this dependence underpins any special teleological status for the latter must acknowledge that the purpose of at least some operations of biological self-preservation, such as metabolic repair and regeneration, cannot be reduced to information processing and syntactical transformations alone.

This is not a problem if we take the ESIA cycle instrumentally as a useful tool for modeling specific behavioral phenomena, such as the regulation of body temperature. But if our model is supposed to provide the basis for a general theory of life, as Friston (2013, 2019) presents the FEP, then to acknowledge that it, like all models, is partial and distorted is not sufficient. The task of a model of "life in general" is to highlight the *right* things and neglect only those contingent features of the particular instances we happened to have encountered.

That all living systems we know of are metabolic systems, and there are arguably no naturally arising nonliving metabolisms, is a reason for thinking metabolism might be important, but not a conclusive one. All living systems that we have encountered are also made up of a specific set of amino acids, but to take these to be essential features of life would be chemically chauvinistic, unless there is a principled reason to claim that *only these specific amino acids* are capable of realizing some general property that could reliably distinguish life from non-life in, say, some potential astrobiological encounter.

There is good reason for thinking that metabolism should be a criterion in such a test, quite independent of any putative connections to a bioenactive account of intentionality and agency. Metabolism, properly understood, is not merely an additional and disconnected capacity of an organism, nor does it just mean something that "depends upon energy" as could also be said of a computer game. In the strong sense, as Boden (1999) argues it should be understood, metabolism describes a different mode of existence from that of a machine. The difference is between something whose physical body is constituted via its own activity and an object that persists independently of its own doing.

Matter, in general, prefers to occupy a low-energy configuration. The form of atoms and the locations of electrons in their orbits is, as Bickhard (2009) notes, the paradigm example of how this tendency shapes its organization. In such equilibrium, things and their organization will remain stable as long as the ambient energy is not sufficiently high to destabilize and destroy their cohesion. This makes for good mechanical parts. In a living system, however, things are exactly reversed. Biological components often occupy inherently unstable configurations, and the continuous input of energy is needed to preserve them at this non-equilibrium steady state (Schrödinger, 1951; Bertalanffy, 1968; Nicolis & Prigogine, 1977; Fleischaker, 1988; Juarrero, 1999; Christensen & Hooker, 2000; Christensen & Bickhard, 2002; Collier, 2004, 2008). No engineer would select such parts to build her machines.

In metabolic systems, the relationship between the structure of a system and its activity, or between the constraints it embodies and the dynamics they produce, is fundamentally different from in machines. To a machine, the flow of energy that facilitates its operation is a threat. Inefficiency in how the machine channels this energy allows it to gradually degrade the machine's otherwise stable parts. In a living system, these structural parts—membranes, enzymes, and other parts—are inherently more unstable than the overall system. Unlike in an atom, the stability that they do appear to have is not intrinsic to their internal configuration but is reciprocally dependent on the activity they enable, which secures and channel the matter and energy that is necessary for the continued replenishment that provides this contingent stability (Montévil & Mossio, 2015; Mossio & Moreno, 2015).

It's important to be clear that what I'm claiming differentiates organic parts here is not just that they are intrinsically unstable, but that the particular kind of stability that they do have is extrinsic to that part. They are stable only because of interaction with the broader network of the organism and how the instability of some of these parts fuels the activity of other parts to synthesize their replenishment—thereby giving them a contingent, mutually interdependent stability. In contrast, though radioactive materials are also unstable and continuously decaying, the partial stability that a Uranium-238 isotope has is merely due to the fragile balance of fundamental forces at the subatomic level *within* that atom—there need be no additional inflow of energy or matter to sustain that stability, and its breakdown does not power any process that replenishes it.

So, organisms are not merely homeostatic mechanisms that remain stable *through* change; their stability is *dependent* on change. As Jonas (1953) criticized the cybernetic approach that underlies the free energy framework, “A feedback mechanism may be going, or may be at rest: in either state the machine exists. The organism has to keep going, because to be going is its very existence” (p. 12) With my bioenactivist hat on, I, like Jonas, would like to describe this in terms of the intentional directedness of the organism toward its environment as a source of matter and energy required to achieve the immanent teleology of its own continued existence. But for the avoidance of controversy, we can just call it the dependence of a non-equilibrium structure on the continuous flow of energy required to sustain it. The fact that the existence of an organic structure *depends* on material turnover, as opposed to merely allowing for it, is a difference between organic and nonorganic existence that our theory of living systems needs to recognize—even if we have no interest in the bioenactive project of interpreting this in intentional terms.

The FEP’s fixed statistical/causal network and its attendant Markov blanket describe only how a system’s structure constrains its dynamics; they do not address any reciprocal dependence of this structure on those dynamics in turn. Just as a steam engine does not need to be in constant operation to continue to exist, and just as a laptop may be turned on and off again with no deleterious consequences to its future capacity for computation, so once Huygens’s coupled pendulums wind down, the connecting beam remains as a constraint on possible interactions should they be perturbed again. Active or not, the pendulums still meet the limitations on potential interactions required for an ESIA cycle, still achieve a free energy minimum, and still possess the beam between them as a Markov blanket.

Nicholson (2018) takes this presumption that an organism can be separated into an invariant structure and dynamical behavior to be the fundamental error of what he terms the “machine concept of the organism,” and one nicely captured in the common analogy of food to fuel. As he puts it,

> The problem of equating fuel with food is that it drastically underestimates the physiological pervasiveness of metabolism. No matter how dynamic a functioning machine may be, it is always possible to distinguish the machine’s physical

> frame—which remains fixed—from the materials that flow through it. The actual structure of the machine does not itself take part in the chemical transformations that the fuel undergoes as it passes through it. Instead, it serves as a channel that facilitates the exchange of materials as fuel is converted into waste. An organism, in contrast, changes wholly and continuously as a result of its metabolizing activity . . . This is why the fuel-food analogy is so misleading, and why the stability of a machine—despite its apparent dynamicity—ultimately resides in an unchanging material structure. In machines there is a specific "inflow" and a specific "outflow." In organisms everything flows. (Nicholson, 2018, p. 145)

It is trivial to note that all machines depend on energy to operate, and we could easily create a machine whose operation is dedicated to the harvesting of it—a sun-tracking solar panel would suffice. But the solar panel's structure does not depend on its success. It need not store energy to make it through a foggy day and it will not disintegrate if it runs out.

What explains the persistent tendency to abstract away from the dependence on material turnover metabolic systems in theories of the living? Boden (1999) argues this tendency may stem not so much from a principled position as to its irrelevance, but rather from the desire to separate logical form entirely from material instantiation and identify the "essence of life" with the former, such that we might hope to simulate it, or even create it, in a virtual medium-independent way. Where other proposed criteria of life, such as growth or adaptation, do not make explicit reference to flows of matter and energy, metabolism is defined in terms of these, and so cannot be straightforwardly captured in purely syntactic or informational terms:

> The only reason for proposing that we drop metabolism from our concept of life is to allow a strictly functionalist-informational account of life in general, and A-Life (artificial life) in particular. The same applies in respect of suggestions that we weaken the notion of metabolism . . . and substitute mere energy dependency (with or without individual energy packets). The only purpose of this recommendation is to allow virtual beings, which have physical existence but no body, to count as life. These question-begging proposals have no independent grounds to buttress them. (p. 8)

The desire of the A-life research program of the 1980s and 1990s to be able to create life in a computer that Boden criticizes is one motivation for this focus on the formal properties of living systems to the exclusion of all else. Such a move may also stem from a more general desire to take formal models to be exhaustive of the phenomena they are supposed to describe—as when scientists become so seduced by the success of their models in serving some particular purpose that they overlook all that they cannot explain and reflexively incorporate arbitrary limitations of their model into their very concept of the target phenomena. This, as Merleau-Ponty (2004/1961, p. 292) argued, was the foundational mistake of the "absolute artificialism" of cyberneticians, like Ashby (1962), who claimed to be able to explain everything from the economy to the behavior of individuals in terms of feedback control.

As Chirimuuta (2020) describes,

> The "absolute artificialism" of cybernetics is a kind of vicious circularity: the cyberneticist has an understanding of organisms based on selective attention to analogies with machines and then uses this conception of organism to inspire the building of new devices, which are then projected back onto living organisms as models of their workings, and through the cumulative and recurrent effect of this process it becomes impossible to think of the organism—including the human being—in any other terms than as a tool, a thing to be manipulated and an instrument at the service of interminable projects of intervention and control. (p. 450)

I take the development of the FEP to have fallen into a similar trap, motivated by a belief in the priority and exhaustivity of mathematical formalization, and a desire to be able to use this alone to "answer the questions traditionally posed by metaphysics; i.e., what does it mean to be a thing that exists, what is existence, etc" (Ramstead et al., 2021, S.43).

That metabolism is not a purely formal phenomenon does not preclude any of its properties being *described* by a model, nor does it mean that it can only be realized by the specific set of amino acids that makes up the metabolisms we are familiar with, nor even that it can only be instantiated at the level of molecular synthesis. In chapter 10, I will discuss Moreno and Mossio's (2015) attempt to describe the functional organization of metabolism in terms of constraint closure. Their account, I will argue, captures the specific way in which flows of matter and energy must be organized so as to realize a metabolic network, thereby placing tight constraints on what could realize such a thermodynamic organization, without arbitrarily tying this to one particular set of chemical components.

The crucial point here is that we can give a multiply realizable functional specification of some phenomena without treating this phenomenon as medium-independent to the degree that a formal operation is (Piccinini, 2020). So long as a medium can realize the relevant syntactic operations that define a particular operation, then anything could be used to realize addition or integration, whether water in tubes or electricity in wires. In contrast, when it comes to splitting an atom or seasoning a chicken, the material matters. A neutron or a proton will both do well for the former task, but an electron will not. For the latter, parsley, sage, rosemary, or thyme are all good options, but strychnine is not.

In the case of metabolism, the constraints on possible realization are derived from energy management requirements—namely the need for a tight synchronization of energy flows between the exergonic release of energy from the breakdown of complex molecules and its endergonic uptake in the synthesis of new ones. These timing constraints are fixed by the reactivity of the specific components making up a particular cell and, with thousands of reactions making up the metabolic network of a single cell, this web of constraints is tight indeed. As Ruiz-Miazo and Moreno (1999) argue, when approaching life from the perspective of metabolism, one is forced to consider how a description of living systems in terms of an "abstract-relational logic" can be "geared in with the implementation or physical realization of some effective management of those energy flows" (p. 46).

So, the reason a molecule like adenosine triphosphate (ATP) is able to play the role of an energy currency, between these exergonic and endergonic reactions, is intimately tied to its solubility, reactivity, and energy density. Alternative energy currencies are available, but the range of possible chemical realizers for this role is extremely limited.

A crucial consequence of the connection between metabolism and medium-dependent energy management concerns is the creation of a distinction between an exact simulation and a genuine realization. Such a difference is not available for a purely formal operation. There is no difference between a perfect simulation of multiplication and its genuine realization. One might say that a particular machine only approximates multiplication, but any non-approximate procedure with the correct formal properties will literally *be* a multiplier. If something is medium-dependent, in contrast, then a simulation in the incorrect medium, no matter how detailed, will remain merely a simulateon because the medium in which it is realized will not instantiate the necessary non-formal properties. The difference between the transistors and diodes that instantiate your Sim's house and a real building is not one of approximateness or partiality. No matter how detailed you make your simulation these transistors and pixels will never provide shelter from a storm.

So medium independence is crucial to the hope of strong artificial life, as stated by Langton (1989), "to build models that are so lifelike that they would cease to be models of life and become examples of life themselves" (p. 63). It is also crucial to the FEP's hope to characterize "everything of interest" about life in purely formal terms. But neither Frankensteinian aspirations nor a mania for mathematical models are good reasons for rejecting the importance of something with such significant consequences for an organism's existence as its metabolism. If metabolism is important, then models of life will always remain mere models, and descriptions of their necessary properties will not automatically carry over to the living systems they are supposed to describe.

When we look at an organism through the lenses of a machine-substantialist, we abstract away from material turnover to treat parts as stable. Such a perspective blankets over how the extrinsic and contingent stability of the enzyme sets it apart from the intrinsic stability of the atom, concealing its different costs and consequences. As I will argue in chapter 10, these differences are vital for providing a naturalistic account of intentionality and purposiveness as a property of organic systems alone.

Still, all this is not to say that we cannot occasionally treat organisms like machines and give mechanistic models of them in cases where we can argue that this distinction is irrelevant. Abstracting away from this fluidity may still be justified by how it allows us to apply the FEP's second requirement, namely the stability of the probability distribution over the possible states of the system. Even if this is not a first principle from which "everything of interest about life" can be derived, it may, like natural selection, describe a regularity that living things in general tend to conform with, and by means of which we can predict *some* of their behaviors.

8.3 From Stability of Realization to Stability of Organization

When describing organisms as machines, we treat them as things with stable parts that are individuated not by their material basis but their role in a fixed organization. This machine-substance concept of the organism is distinct from the naive concept of a material-substance, but, as Dupré and Nicholson (2018) describe, is equally rejected by proccessualist views. In analyzing this, Nicholson (2012) quotes Glennan's (2002) definition of a mechanism as "a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (Glennan, 2002, p. S344).

This definition has two criteria: decomposability and invariance. The former of these, the claim that particular phenomena can be explained in terms of local interactions between distinct parts, is most associated with the "new mechanist" movement currently ascendant across philosophy of science (Bechtel & Abrahamsen, 2005; Machamer et al., 2000; Glennan, 2002). It is this claim of decomposability that formed a central target of the organicist movement in early twentieth-century biology, a precursor and close relation to contemporary, processualist accounts of the organism (Dupré & Nicholson, 2018; Peterson, 2017; Chirimuuta, 2020). Living systems, the organicists argued, are not decomposable mechanisms but "organic wholes" with emergent behavior that alters the activity of their parts and cannot be reduced to local interactions between these.

I agree with the organicists that the way in which the parts of organisms are mutually dependent on one another for their contingent stability makes decomposing them for explanatory purposes, as mechanistic accounts do, more of a distortion than in the case of a machine, where we can literally separate out the fuel injector and the intake manifold from the engine. I discussed this in the previous section, however, and my concern here is not with exploring the consequences of this for the legitimacy, or not, of mechanistic explanation.

Rather, with respect to how this machine concept of the organism influences the FEP, the relevant issue is the presumption that the organism itself (and not just some approximate model of it) is reducible to an invariant set of equations describing its behavior. If this is the case, then we could identify some invariant formal parts in terms of the variables of these equations—even if those parts do not correspond to discrete things that can actually be separated from the organism in reality.

It is these two essential invariants of a machine-substance ontology that are expressed in the two stability requirements making up the FEP's definition of a "thing." So if, for any living system, we can identify a fixed set of formal parts whose dynamics can be described by some invariant rule, then it will qualify as a machine in the sense I am using the word. If this invariant rule is one of convergence to a steady state (or ergodicity), then it will be, more specifically, a free energy minimizing machine.

I have already touched on the implausibility of thinking that a drive toward a steady state might serve as a first principle from which we could derive everything of interest about the organism. Yet, for all that this is a particularly impoverished framework for characterizing life, I want to argue that rejecting it as the "wrong sort of invariant" does not go far enough. In examining how the FEP's invariants fail to track the identity of an organism over time, I believe we will discover the impossibility of identifying any invariant rule that can do so.

As we've seen, Friston presents the invariant homeostatic rule that defines our steady-state system in two different ways: (1) as a fixed joint probability distribution over states that must be preserved, described as the "generative model," and (2) as a fixed stochastic differential equation that defines the random dynamical attractor of a steady-state system. While the former version is more common in philosophical treatments of the FEP as a Bayesian brain theory, this statistical description is derived from the dynamical requirement.

So, *if* every "thing" is indeed a random dynamical system with a point, or limit cycle attractor, then Friston's claim that its existence over time is dependent on its states changing in conformity with the steady-state equation necessarily follows. The latter is not an *a priori* truth about all existence, however, but is derived from a perfectly defeasible antecedent. As I will describe, as soon as we define something in terms of dynamical system theory, we are committing to the substantialist ontology of essential and invariant rules that makes this thing what it is. The crucial prior question that work on the FEP has overlooked is whether dynamical systems theory in general, and a random attractor in particular, are the appropriate conditions for defining an organism.

A dynamical system consists of a set of equations of motion and a fixed phase space of dependent variables that capture all the possible states for that system—including "states of motion," such as velocity. So, for the simple case of a frictionless pendulum, we have length, mass, and gravitational force as parameters, a two-dimensional phase space of angular position and velocity as dependent variables, and a differential equation relating these to determine possible trajectories through that phase space.

As Longo et al. (2012) describe, these equations and the phase space are standardly derived from what physicists and mathematicians call the "symmetries" of the system, invariants that must be continuously preserved throughout any transformation it undergoes. In the case of our idealized, frictionless pendulum, this is the overall energy, which remains constant while potential energy (determined by position) and kinetic energy (determined by velocity) change. The equations of motion specify these invariant-preserving transformations, while the phase space consists of all and only those complexes of states that can occur within some invariant preserving trajectory (see figure 8.1).

In its focus on continuous change over time, dynamical systems theory has been touted as offering an account of cognition that could supplement the discrete, atemporal symbol manipulation of Turing-style computationalism (Van Gelder, 1995; Chemero, 2011). Moreover, in allowing our phase space to be made up of collective variables, which need not correspond to the distinct "real parts" of a compositional mechanism, dynamical systems theory has been embraced by those in the enactive and embodied tradition of cognitive science, seeking to describe the putatively holistic and emergent behaviors of complex systems (Chemero & Silberstein, 2008; Stepp et al., 2011; Batterman & Rice, 2014; Ross, 2015).

There is some debate whether the covering law descriptions provided by DST (Walmsley, 2008) qualify as explanations, and whether they offer a genuine alternative to mechanistic analysis (Kaplan, 2017; Kaplan & Bechtel, 2011; Kaplan & Craver, 2011). Irrespective of whether DST does indeed provide an alternative, holistic explanatory paradigm, insofar as it

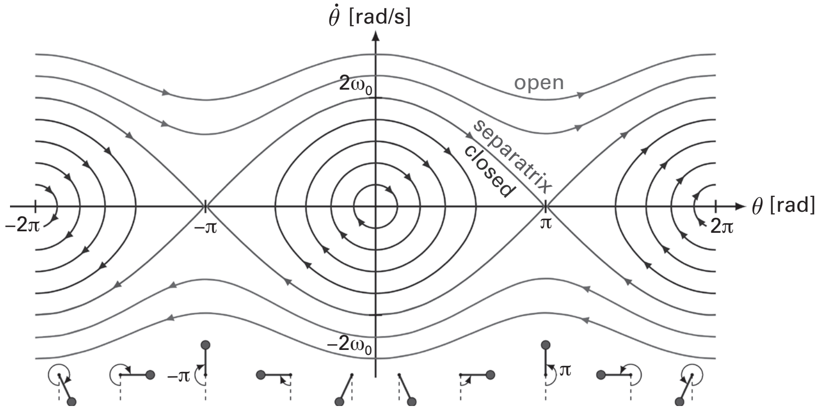


Figure 8.1
The phase space for a simple frictionless pendulum (copyright TikZ.net).

defines a system in terms of a set of invariant equations and parameters that dictate its dynamics and are not affected by them in turn, it nonetheless retains the substantialist commitment of the machine picture that concerns us here.

As Koutroufinis (2017) puts it,

> There are two essential elements of Turing machine-logic. First, sharp distinction between state variables (what is calculated) on the one hand and parameters, independent variables and operators (what is given from outside) on the other. Second, no generation of new general types of intrinsic causal factors, and thus restriction to a fixed state space. (p. 34)

In its dependence on the assumption of some invariant transformation rules with a fixed and finite space of possible states, dynamical systems tools allow us to describe a target system only insofar as it treats this target as following the invariant logic of a machine. It may offer an alternative language to computational or mechanistic analysis, but its descriptive scope is equally constrained.

This doesn't mean that dynamical systems models require us to commit to a substantialist ontology regarding the things that our equations are supposed to be models of. If we take "the system" for which these properties must remain invariant to characterize descriptive content of our particular approximate model, rather as a real-world target, then the question is only whether our model has enough flexibility to characterize all of the behaviors of our target that interest us. Similarly, to say that free energy minimization and the active inference framework provide a useful set of tools for modeling some behaviors of a system does not commit one to treating free energy minimization as an existential imperative that defines this particular system.

If, however, we take this model to be a *definition* of what it is to be, say, a pendulum or a person, then we commit ourselves to a machine-substance ontology where these extrinsic factors of parameters, phase space, and equations constitute the essential features of the particular "thing" in question. If they change because, say, the string of the pendulum breaks, then we say that particular "thing" has ceased to exist. It is this commitment to what I am calling the machine-substance account of the organism that forms the ontological backdrop of the FEP as a theory of things. Applied to a pendulum, it seems reasonable enough. But how well does the FEP in particular, or dynamical systems theory and the machine-substance view more generally, serve for identifying a necessary feature of organismic existence?

9 Seeking Stability with the FEP

We have seen that the steady-state equation of the FEP allows us to describe a system as changing in two ways that preserve a stationary probability distribution. The first kind of change is a temporary fluctuation away from an attracting set of states—so long as the duration spent in these perturbed states remains consistently small. This is guaranteed by the dissipative component of the ESIA-cycle equations, which was stipulated to exactly counterbalance these fluctuations. The FEP can also allow for systems that cycle around equally likely states, described by the (optional) solenoidal component of the ESIA-cycle equations. This has been proposed as a distinguishing feature of biological systems but, as mentioned in chapter 4, seems inadequate to the task given that plenty of things, from pendulums to planets, can cycle through recurring sets of states for an extended duration before gravity and friction eventually bring them to a halt—yet we are not inclined to consider them alive at any point of their doing so.

As a result, the literature tends to presume that all that is important to identifying these systems is the single random dynamical attractor, the single free energy minima to which it will eventually converge and remain within. Ramstead et al. (2021) succinctly express this.

> On this account, to exist as a living system entails continually revisiting the neighbourhood of the same set of states (e.g., remaining within a certain range of body temperatures or ecological environments). Technically, such systems are endowed with a random dynamical or pullback attractor. This engenders a non-equilibrium steady-state density that we can associate with the phenotype of a living system. (p. 110)

And, as eight coauthors assert in a more recent paper this steady-state density and consequent surprisal over states

> can be cast as the ontological potential of a system (when it exists). We define an ontological potential as an abstract potential that induces an attractor for the dynamics of some system. It is ontological in the sense that it characterizes what it is to be the kind of system being considered. This is simply because the system is attracted to sets of states or paths that are characteristic the kind of system that it is, by definition (since they are attractor regions of that system). (Ramstead et al., 2023. p. 10)

This is a particularly simple characterization that elides the wide variety of itinerant trajectories that can potentially be described in dynamical systems models, as Friston (2019a) acknowledges when he states that

> symmetry breaking (i.e., divergence of nearby trajectories to different regimes of phase-space) is a hallmark of non-equilibrium dynamics (Evans & Searles, 2002) and is intimately related to phenomena like self-organised criticality in dynamical systems (Bak et al., 1988; Vespignani & Zapperi, 1998). Indeed, much of complexity science addresses the problem of how to formalise multiscale, itinerant and chaotic dynamics . . . *In this monograph, we will elude many of the finer details (and phenomena such as bifurcations, frustration and phase transitions) and suppose that the interesting behaviour of self-organising systems can be captured by non-equilibrium steady state densities with the right sort of shape.* (p. 20) [my emphasis]

In characterizing lifeforms, this focus on systems that repeatedly cycle through to the same region of phase space has some plausibility, as the continuation of an approach that treats homeostasis as the fundamental principle of organisms, which we saw in Ashby. That logic, Ashby and Friston both argue, extends beyond bodily regulation to every activity of even complex creatures like ourselves: from the bodily rhythms of our respiratory, cardiac, circadian, or hormonal cycles to our daily routines, weekly schedules, and annual festivals. A nice example of this is at the behavioral scale comes from a study on the records of millions of mobile phone users that found that over the space of three months, their location could be predicted with 93 percent accuracy. irrespective of how far the individuals tended to travel (Song et al., 2010).

Still, the observation that organisms often regularly revisit the same states is not proof that they *have* to as a condition of continued existence. A pervasive tendency is not the same as a necessity. Humans may often be creatures of habit and routine, but they are also creatures that undergo dramatic identity crises. They may regularly revisit the same locations over some three-month period, but at the end of those three months, they may leave and never come back. A small-town girl with dreams of stardom ventures off for the bright lights of the big city; a worn-out researcher ups sticks for a log cabin in Siberia, and a certain type of man reaches a certain age and trades in his daily commute and his VW Golf for a motorboat and a fishing license.

Perhaps the best-known objection to the FEP then, often discussed under the heading of "the dark room problem" (Sun & Firestone, 2020; Friston et al., 2012), is the point that a drive to minimize free energy fails to even touch upon this diversity of creative and novelty-seeking activities that are characteristic of human behavior.

A common response to this objection is to attempt to conceptualize such behaviors as temporary excursions that increase free energy in the short term, in service of increasing the agent's representational and behavioral repertoire—thus better positioning it to reduce free energy over the long term (Seth et al., 2020). Still, while a drive to curiosity may serve surprisal minimization over the long term, the solution to this exploit-explore trade-off is not something that can be directly derived from the imperative to reduce surprisal with regard to a particular state and a single generative model (Millidge et al., 2021b). For this reason, describing this explorative behavior requires the construction of a new quantity, such as "expected free energy." Unlike variational free energy, this is scored over possible trajectories of a system, rather than with regard to individual states but, as Millidge et al. (2021b) describe, it is not just variational free energy extended into the future but a different functional altogether. While minimizing variational free energy involves minimizing a complexity term capturing revisions to the agent's model, in expected free energy, the equivalent of this complexity term is subtracted rather than added. This means that, when minimizing expected free energy, changing your model to incorporate new information is now *rewarded*, rather than penalized. While this is fine for the purposes of active inference models, as Millidge et al. (2021b) describe, this curiosity motivating component is not a principled derivation from the surprisal-bounding requirement of the FEP.

To save the FEP's foundational status and its grounding in an imperative of surprisal minimization, we need a means for allowing a system to persist through more than just temporary excursions from a single steady-state regime—a way for it to preserve its identity while moving *between* stable regimes. There are three ways this has been done in the literature. The first, which I will call the "allostatic strategy," involves limiting the scope of the FEP's stability imperative to a privileged subset of "essential variables," thereby freeing all the other organisms' states and behaviors to change in whatever way serves this restricted homeostatic task.

In a similar manner, the second "hierarchical" or "multiscale" strategy also involves dividing up our system. Yet, where the allostatic strategy introduces a divide between types of variables, the "multiscale" strategy instead draws a distinction in terms of timescales, proposing that what appears to be a change from one steady-state regime to another at one timescale, can be conceptualized in terms of a higher-level stability that operates over a longer timescale. Just as active inference described learning in terms of a hierarchy of generative models in the brain, so too might we describe adaptive changes in the behavior of an individual organism in terms of a hierarchy of models, realized at different levels of biological organization.

I will argue that neither strategy is successful; organisms are constrained by neither the statistical stability of privileged physiological variables nor of higher-order parameters. As such, advocates of the FEP have recently proposed a rather surprising third strategy that, they claim, involves abandoning the requirement of statistically stable steady states (Sakthivadivel, 2022; Ramstead et al., 2023; Friston et al., 2023). To abandon such a requirement, however, would be to render the FEP a vacuous tautology that places no meaningful constraints on the systems that would qualify as complying with it.

The difficulty with securing any level of description of biological systems at which the FEP necessarily and meaningfully holds does not just reflect the limitations of this account, however. As I will argue in section 9.5, it is the product of deeper issues with the substantialist attempt to define the identity of an organism in terms of *any* sort of stable invariant. Organisms are more like processes than substances, and their continuity is derived from how each stage of their development relates to the previous one, not from any invariant feature that persists throughout such changes.

9.1 The Allostatic Strategy

9.1.1 From Homeostasis to Allostasis

It's becoming clear that it is impossible to maintain the steady-state requirement of the FEP, its maximally comprehensive form as a foundational principle that directly restricts *everything* an organism must do to survive.

What we need is a modification that will allow this principle to incorporate the possibility of a system's persistence throughout the kind of changes that amount to more than just spinning around on the spot.

The most natural move here is a familiar one from the tradition on biological regulation—to distinguish between homeostatic and allostatic mechanisms where the former involves only those "systems that are truly essential for life"—for instance, systems governing the regulation of physiological variables, such as body temperature or blood oxygenation (McEwen, 2000, p. 173). In contrast, the notion of allostasis instead refers to second-order mechanisms, governing proxy variables and behaviors whose stability is not directly connected to survival. As such, the key feature of allostasis is that it can encompass various forms of change beyond simply the preservation of an invariant steady state.

To demarcate allostasis purely in contrast to homeostasis still leaves open a variety of positive proposals about how this more flexible form of regulation is managed and how its target should be conceived. Accordingly, as Corcoran and Hohwy (2018) describe, a number of ambiguous and contradictory concepts of regulation have all been advanced under the label of "allostasis" since the term was first introduced by Sterling and Eyer (1988). In the context of the free energy principle specifically, the most recent treatments characterize allostasis as a distinctively anticipatory and model-based way of selecting a trajectory, rather than a single set of states, for the regulated variable (Barrett et al., 2016; Barrett, 2017; Pezzulo et al., 2015). Because many disrupters to homeostasis are periodic and regular, so an allostatic model allows the body to act to minimize expected free energy with respect to a predicted trajectory of states, and thus to prevent a disruption to homeostasis *before* it actually occurs. To take one example, rather than waiting until the onset of hypoglycemia before taking compensatory action, animals will anticipate that glucose levels will decline and thus seek out and eat a meal in advance of such an occurrence.

So, at a technical level, the active inference framework has the tools to model both allostatic and homeostatic regulation mechanisms, once the steady-state ESIA cycle of the FEP is constrained to describing the homeostasis of essential variables alone. But this ability to assimilate predictive regulatory mechanisms is not the crux of the issue that allostasis raises the FEP. The central problem is how the FEP forces us to conceptualize the *target* of allostatic regulation. If allostatic regulation is to be derivable from the FEP's first principles analysis of survival as surprisal minimization, then allostasis and homeostasis cannot just be treated as two distinct regulatory modes, each with their own independent logic. Rather, the former must ultimately be in the service of the latter—that is, in the service of preserving the stability of some set of "essential" variables. The necessary stability of these variables is essential if the steady-state formulation of the FEP is to retain its purchase on capturing the fundamental imperative of the organism. As such, most work on allostasis in the free energy literature explicitly describes the purpose of allostasis as being in the service of a "fundamental imperative" toward homeostasis (Gu & FitzGerald, 2014; Pezzulo et al., 2015; Seth, 2015).

The issue with maintaining a commitment to some limited form of homeostasis is not dissolved by merely introducing the possibility of a distinction between essential and proxy variables. What we need is a principled basis for distinguishing this essential subset of variables whose stability defines the organism's continued existence. The FEP itself does not provide a means to derive this distinction. As described in chapters 5 and 6, neither the Markov Blanket nor the ESIA cycle provides a principled basis for dividing what constitutes an organism from what does not. Once we have used other means to select our relevant variables, *then* we can perfectly well ask if their tendencies and dependencies allow them to be described in terms of Markov blankets, ESIA cycles, and steady states, but these tools do not provide our initial demarcation, only some constraints on the allowable interactions between whatever variables we might select.

The identification of what is essential to an organism's survival, and what is not, seems like the first thing we would want a first principles analysis of life to provide. Moreover, we should expect this principle to do more than simply look at which states were in fact kept stable through an organism's life and then describe those as the states that *must* have been kept stable. Insofar as we accept global determinism, this may be trivially true, but it cannot address any distinction between the essentiality of blood-glucose regulation versus the essentiality of hair color.

Even if the FEP does not identify our essential variables for us, this would only undermine its status as a first principle, not the applicability of free energy minimization as a modeling framework. If we have a prior independent means of identifying our essential variables, then describing these in terms of free energy minimization may amount to no more than a statistical redescription of something we already knew. Still, quantifying their homeostatic stability in terms of the surprisal of certain states would provide a foothold for the active inference framework and, potentially, for fruitfully modeling second-order allostatic regulation in terms of expected free energy minimization.

This approach will only work, however, if we grant the assumed premise that the organism actually has any essential variables at all. Insofar as we treat allostasis as being in the service of homeostasis, all this version of allostatic strategy does is permit a reduction in the scope of which variables are considered essential. The FEP's treatment of allostasis still retains the requirement that there be some essential variables and, thus, remains committed to the idea of organisms as substances, identified by some invariant features that must be preserved over time.

Allowing that some of the states or behaviors that constitute an organism are not essential to its identity may render claims of invariance more plausible. Still this conceptualization of allostasis as a mechanism "in the service of homeostasis" runs in direct opposition to Sterling's proclaimed intention in introducing the concept specifically as a means "to replace homeostasis as the core model of physiological regulation" with a view of regulation as being directed toward meeting the demands of fitness, that is, toward survival and reproduction (Sterling, 2004). Crucial to Sterling's point is that such demands simply cannot be reduced to stability through fluctuation of *any* physiological variables. As he describes in the case of a paradigm "basic" physiological variable, such as arterial blood pressure,

> In correlation with identified external stimuli and mental states, it varies more extremely. As the subject dozes in lecture, pressure falls to 80/50. When he is jabbed with a pin, pressure spikes to 150/70; then, having recognized the joke, he again relaxes, and the pressure sinks to 80/50. During sexual intercourse, pressure spikes to 170/90 and then falls profoundly during sleep to ~70/40 with one hour as low as 55/30. In the morning pressure steps up nearly to its level during sex and remains high for hours. (p. 7)

The fluctuations from stable state described here are neither regular cycles nor random perturbations to be countered but adaptive responses to specific events that serve vital context-sensitive functions. In the case of blood pressure, this would be the delivery of increased oxygen to the cell as and when it is needed to support an increased metabolic rate. What determines the adaptiveness of this change is determined not by *how regularly it happens* but precisely by *when* it happens—by the precise circumstances under which it occurs. If my blood pressure does not increase when I am faced with a 500-pound grizzly bear, the cause of my death will not be statistical. I will die due to lacking the necessary oxygen required to power the energetic escape required in this unusual situation. In contrast, if blood pressure is regularly high in situations where no energetic activity is needed, then the mere fact that this is my usual state does not make it a biological good that I am constantly expending more energy than needed.

This same story, Sterling argues, holds true for essentially all physiological parameters that we consider vital. "All fluctuate with different amplitudes and time constants, and these fluctuations all share a single goal. Yet the goal is not constancy, but coordinated variation to optimize performance at the least cost" (p. 9). The regularity that we do observe, he argues, is more appropriately conceptualized as the result of contingent regularities of demand (i.e., the fact that the average academic rarely finds herself being chased by 500-pound grizzly bears) rather than as being a necessary principle of survival in its own right.

We can find this context-sensitivity in the parameters of survival in all organisms, even the simplest single-celled lifeforms. A classic example is the lac operon mechanism of the *E. coli* bacterium. Ordinarily, such bacterium may be well characterized as steady-state systems, cycling through the metabolization of glucose molecules, with some fluctuation around a stable intake level. Yet, when glucose levels drop and lactose levels rise, a coding region in the *E. coli*'s DNA is activated, triggering the production of enzymes to allow for the metabolization of lactose as well. This amounts to a move from a steady state that was describable by a stable probability distribution with zero likelihood of lactose metabolism, to a new steady state where this now has a high probability.

If we take the free energy principle as applying to the state of such low-level physiological variables, such as blood pressure or rate of lactose metabolism, if we accept its definition of existence in terms of the stability of such variables and death as a phase transition away from steady state, then we are compelled to describe this transition away from a stable pattern of lactose metabolism as the *E. coli*'s death.

9.2 Multiscale Stability

If the needs and demands of the organism, and thus the performance of regulatory processes that serve them, cannot be grounded in the constancy of *any* particular lower-level physiological variables, then simply introducing a second-order allostatic process in the service of selective stability will not solve our problem. Still, at this point, the homeostasis-defenders may object that we still have not shown that such needs cannot at least be described, even if not derived, in terms of stability of variables. One can always suggest that there must still ultimately be some higher-order variable, whose stability captures the notion of fitness or performance, and which governs all these shifting context-sensitive changes in other variables (Stephan et al., 2016). Corcoran and Hohwy (2018), for instance, argue that all Sterling's example shows is that blood pressure is the wrong variable to focus on. Instead of taking higher-level regulatory changes to be in the service of lower-level invariance, then we might instead take all this lower-level change to be in the service of a higher-order stability.

The appeal to more complex, multistable attractor landscapes is a common way to incorporate what looks like a breakdown of a stable regime. In the context of modeling neural dynamics, for instance, a pre-FEP Friston (2000) notes that a change in stable regime can be described in two ways: (type 1) as a movement between a set of invariant metastable attractors determined by an unchanging equation, or (type 2) as a change in the control parameters of the equation governing the dynamics of a system and thus a change in its attractor manifold. If type 2 change were irreducible, that is to say, if the alterations of the transition equation parameters were not derivable from some other equation, then the potential behavior of said system would outstrip the formalizations not only of the FEP but of *any* mathematical model.

Yet, as Friston argues, type 2 changes may also be described by simply incorporating an equation governing how these control parameters change into the overall set of equations and state space describing the behavior of our system. So long as said equation does not itself have further changing control parameters, then this procedure allows us to turn a system that appears to undergo type 2 changes into a more comprehensive system that only undergoes type 1 change, as it transitions through a fixed, multistable attractor manifold. This, Friston argues, is

> a crucial observation because it suggests that the distinction between type I and type II complexity is simply a matter of perspective. In other words, what may be a type II complexity from the point of view of one system may turn out to be a type I complexity when one "stands back" and considers a larger system in which the first was embedded. (Friston, 2000)

Thus, we might defend the free energy principle by arguing that all we have shown is that the survival of an organism does not depend on the preservation of a stable probability distribution over physiological variables directly, but rather on the of the stability of some higher-order set of parameters determining a recurring cycle between different steady states. As we saw in section 3.3, this is exactly the strategy that the active inference framework has taken toward the incorporation of learning, where changes in lower-level parameters are described as being in the service of minimizing free energy with respect to a higher-level invariant model (Pezzulo et al., 2018).

If the free energy principle is to maintain its scope as a first principle that governs everything an organism does, then it must *always* be the case that we can identify and incorporate a fixed equation governing all the changes that an organism undergoes, and, specifically, that this equation has the right steady-state form to be translated into a stable generative model for that system.

The problem now, however, is that we can no longer derive our measures of surprisal and our generative model of the organism by just taking the first recurring pattern of behavior we find—for this may just be one temporarily stable phase in the preservation of a higher-order stability, rather than being reflective of that system's full generative model. So, how do we identify these higher-order stabilities?

One early proposal can be found in Friston and Ao (2012) and Hesp et al.'s (2019) suggestion that the multiple scales of free energy minimization are instantiated not only at different timescales in an individual system but rather at different spatial scales beyond the individual—that is to say, at the level of a group or a species. If this is the case, then rather than the post hoc procedure of following an *E. coli* bacterium throughout its entire life, to establish a stable probabilistic model for it, we could instead derive the statistics of its life cycle from a snapshot of an ensemble of *E. coli* at a particular point in time.

Doing so, we can find that for any randomly selected time period, the percentage of bacteria that are in the lactose metabolizing state is constant—say 10 percent. Thus, we can derive a model for the *E. coli* bacterium as having a 10 percent probability of metabolizing lactose. From the perspective of this species-level model, the transition of an individual *E. coli* bacterium from having spent nine days only metabolizing glucose to suddenly starting to metabolize lactose would no longer appear surprising. Instead, it would be a predictable and recurring phase transition under the model for *E. coli* in general.

To justify this translation between a generative model for the "ensemble at a time" and one for the "individual over time" relies on not just the assumption that each system will converge to a steady state, but the stronger requirement of *ergodicity*—such that we can view the probability distribution over states for different iterations of a system as interchangeable, in spite of their differing initial conditions. As discussed in chapter 4, a reason for shedding this for the weaker steady-state requirement is that ergodicity is actually rather difficult to prove outside of idealized models. A particular issue, as Colombo and Palacios (2021) point out, is that the time for the average behavior of an individual to converge to the ensemble average could be arbitrarily long, and far outstrip the duration over which that system exists (see also Palacios, 2018; Gallavotti, 1999).

The trouble with this is perhaps not so obvious for a creature with a comparatively small behavioral repertoire, like an *E. coli*, where the idea that each of its members could be defined by one invariant species-level model might be somewhat plausible. Yet applied to say, human beings, any probability distribution over all the states that members of the species might feasibly occupy would cover such a large variety of possibilities as to be completely uninformative about the dynamics of any individual.

Life is short. Even if there were a fixed attractor manifold for human viability, no single member of our species will ever have time to explore enough of it for their individual dynamics over time to converge to a stable probability distribution. In this respect, the FEP's "species-level" stable probability distribution would not only be an asymptotic idealization that is never actually realized by the kind of systems that we are interested in but, crucially, one that places no informative constraints on the dynamics of these systems over the timescales on which they actually do exist.

9.2.1 The Problem of Development

This timescale issue is one reason for rejecting the idea that we can use an ensemble-level probability distribution to describe the expected behavior of a complex individual entity. Still, the reason ergodicity is especially problematic for living systems is not adequately captured by simply noting that they consist of so many parts and degrees of freedom that the time required to converge to a recurring pattern of behavior would be exceptionally long. As Colombo and Palacios (2021) and DiPaolo et al. (2022) argue, what makes organisms especially resistive to description in terms of any stable probabilistic model is that they are irreducibly historical systems whose activity canalizes their future dynamics and changes their own space of possibilities.

As we develop, as we age, states that were possible in our past not only become further away from our current trajectory but leave our possibility space altogether. No matter how long we hang around, we cannot recapture our youth. In any population of developing systems, there will be various states that have reasonable likelihood at the ensemble level that are no longer possible for certain individuals. It may, for instance, be true of the human ensemble that any individual picked at random has a 20 percent chance of having baby teeth, but this does not mean that there would be a 20 percent chance of my regaining baby teeth if I only live long enough to get around to being a baby again!

Many of the changes that organisms undergo are not part of any larger recurring cycle either within or between steady states. They are irreversible changes like the dramatic transformations of an organism's entire body plan that occur over the development of an undifferentiated bundle of cells into a complex, heterogeneous newborn. Every living system may spend a certain amount of time as a baby, but once adulthood is achieved, neonatal states are simply no longer part of that individual's organism's possibility space, even while they continue to be part of the model for the entire species.

Clearly, accounting for development will require the abandonment of claims such as that "things only exist on timescales over which they are ergodic" (or at steady state) (Friston, 2019b, p. 176). Any timescale that incorporates developmental stages will not satisfy such a requirement, leading to the absurd conclusion that while I may have "existed" from 10 a.m. to 5 p.m. this evening, I did not exist over the timescale ranging from my birth until my first birthday.

If we give up the attempt to justify an equivalency between species and individual-level probability distribution, such that the former could be used to derive the latter, we might nonetheless find a weaker role for the species-level model in determining the behavior of the individual. Without ergodicity, we might still allow that the generative model of the species does describe an attracting set of states that an individual eventually achieves and subsequently maintains—namely a stable adult form. While developmental stages may be part of the stable generative model for the whole species, however, this generative model cannot provide a definition of the individual organism, for whom these are irreversible developmental transitions that will not recur.

So, contra Ramstead et al. (2018) and Friston and Ao (2012), we cannot reduce the existence of an organism to the preservation of a stable generative model. We might suggest, however, that once an organism achieves a stable adult form it will subsequently be describable in terms of active inference and variational free energy minimization with respect to the stability of that final form. This would significantly reduce the scope of the free energy principle from its earlier claim that everything an organism does can be summed up in maintaining the homeostatic stability of its states. Such a reduction in scope is an improvement given the implausible of this initial claim but, unfortunately, the idea that the species (or other biological group) can supply a steady state that every individual of its type will eventually achieve and maintain is not much better.

Unlike the ergodic formulation, this allows room for individual differences during the development phases. Still, if the steady-state model that every individual organism is supposed to achieve is derive from its species, then this still requires that every organism of a type will ultimately reach a state in which they can be treated as interchangeable. Any capacity for individuality is constrained to either the particular path the system took toward its steady state, or to random fluctuations from that steady state.

Moreover, the suggestion that development can be confined to early processes involved in achieving a predetermined configuration, that subsequently stays fixed for the remaining duration of an organism's life span, is exceedingly implausible. It is not only complex humans who undergo midlife crises. Even after having achieved a stable form, many organisms continue to undergo continued, physiologically transformative shifts. Concession is sometimes made in the free energy literature to the metamorphic butterfly as an odd exception to the rule that organisms can be defined by a stationary probability distribution, but little work has been done to reckon with the consequences of this for the aim to "unify all adaptive autopoietic and self-organizing behaviour under one simple imperative; avoid surprises and you will last longer" (Friston, 2012, p. 2).

This kind of phenotypic plasticity is not that uncommon. There are also the 500-odd species of fish known to change sex under the appropriate circumstances (Murata et al., 2021), or the dramatic transition of a placid bunch of short-horned grasshoppers into a swarming, seething gang of locusts, with a quite dramatically different appearance and behavior, when they reach a critical mass (Burrows et al., 2011).

It is not only in development that we find non-recurring transitions. Rather, as DiPaolo et al. (2021) argue, the possibility of such permanent transitions, such as "embryogenesis, life cycle patterns, epigenetic variability, metamorphosis and symbiosis" is pervasive throughout the entirety of an organism's life cycle. This does not mean that there may not be certain states that are contingently preserved throughout these transitions. The point is that we have no issue in identifying a particular butterfly as the continuation of the very same caterpillar that entered the chrysalis, or the adult as the continuation of the baby despite the vastly divergent character of their possibility spaces.

9.3 From Steady States to Predictable Paths

The FEP was initially presented as the claim that everything an organism does can be summed up in the homeostatic imperative to minimize the surprisal of its states. This is transparently implausible when read with maximal scope to apply to every single state of a system, but can be rendered less so if we take it to apply only to some privileged subset of either lower-level physiological states, or to higher-level relationships in terms of how lower-level states change. Still, even when this restricted form of stability is all we're searching for, we've seen how hard it is to identify *any* description of an organism that is both general enough to remain stable throughout all the context-sensitive and developmental changes it undergoes, while still being specific enough describe the survival of that specific organism over time.

Attempting to actually *disprove* that there are any necessary stabilities underlying biological organization feels a bit like scouring the solar system for Bertrand Russell's unperceivable interplanetary teapot. Still, we can do more than throw up our hands and shift the burden for proofs of existence onto the faithful. Just as the artefactual nature of teapots make the possibility of there being one orbiting the sun halfway between Earth and Mars extremely unlikely, so the pervasiveness of irreversible changes at all levels of biological organization should be enough to seriously undermine one's faith in the inevitability of recurrence.

Accordingly, recent work in the FEP has made the rather unexpected change of suggesting that the principle can be formulated without the steady-state requirement that has, thus far, constituted the entirety of its empirical content. Technically, producing this reformulation is relatively simple. We simply swap the requirement that free energy is minimized for a specific state of a system to the requirement that said system must takes a path that minimizes the overall free energy encountered over the course of that path (Sakthivadivel, 2022; Ramstead et al., 2023; Friston et al., 2023).[1] As Ramstead et al. note, this formulation is standard in active inference approaches to action planning, where a system is understood as not only inferring the most likely state that it should occupy but rather how its free energy will change over the course of an entire future trajectory through states. This temporally deep form of inference thus allows it to select the policy that will incur the least expected free energy, integrated over this entire path.

Where previously this anticipatory, or allostatic, process was treated as being in the service of achieving a homeostatic imperative, the above authors now appear to be suggesting that this path-based free energy minimization is the more basic imperative, from which an agent's dynamics can be defined. This, they suggest, would allow its path to incorporate changes in state that do not immediately minimize free energy or surprisal—so long as this increase enables the path to access a greater decrease, such that the total quantity integrated across this path is lower than for any alternative path. To illustrate the idea, compare a law-abiding rollerblader, trundling along a level pavement, to the rebellious skateboarder who temporarily sacrifices some velocity to climb a garden wall, only to gain it back with interest as she flies down the steep basin of the local millionaire's empty swimming pool on the other side. The skateboarder may momentarily decrease her speed, but all in the service of accessing a greater increase in velocity such that the amount she experiences overall is greater than that of the rollerblader.

Insofar as minimizing surprisal over a trajectory does not require that this trajectory ever converge to a particular stable cycle or set of states, so Ramstead et al. (2023) claim that this path-based approach specifically allows the free energy principle to describe historical, nonrecurring processes like development, insofar as it “assumes very little about the dynamics of the system, and in particular, does not assume that a non-equilibrium steady state with a well-defined mode exists” (p. 6).

This apparent lack of content, they argue, does not undermine the FEP but rather reveals that its status is even more foundational and universal than was suggested by its prior presentation as a theory of biological self-organization. Presented in this more general sense, they claim, it can be seen as a version of one of the most fundamental principles in modern physics: namely, Hamilton’s principle of least action. This states that the equation of motion for a physical system can be derived from the requirement that the path it takes between two points will be the one that minimizes the integral of some quantity over the entire path, where this integral is called the “action” of that path. In Lagrangian mechanics, this is defined as kinetic minus potential energy and its minimization underpins the classic parabolic trajectory of a projectile. What makes this least action principle so “fundamental” however is that it can be seen to underpin physical laws in a wide range of areas from the path that light takes through different media, to the behavior of quantum particles. Roughly, we can gloss it as the idea that nature always takes the most efficient path between two points.

Similarly, Ramstead et al. (2023) and Friston et al. (2023) argue that the FEP simply shows how this same principle applies to the dynamics of an inferential process. We simply replace the Lagrangian, composed of kinetic minus potential energy, with the quantity of surprisal (Friston et al., 2023). Insofar as the integral of this over a path must be minimized if the selection of that path is to conform with the requirement of Bayesian inference, so Bayesian inference can be derived from the application of the least action requirement to surprisal. This allows Hamilton’s principle to be extended to the inferential behavior of agents and thus, Ramstead et al. (2023) claim, delivers a “Bayesian *mechanics* of belief.”

Dealing with paths of surprisal certainly expands the capacity of active inference to describe systems and is particularly well suited to when dealing with prospective inference of the consequences of one’s actions. If we also describe the system as minimizing a path of “*expected* free energy,” in which changes to the model are rewarded rather than penalized (as they are under standard variational free energy minimization), then we can describe a much wider variety of trajectories as still being consistent with some sort of free energy principle.

What all this does not change however, is the fact that these quantities and the paths that minimize them are still defined in terms of probabilities or surprisal values. As such, the path-based version of the FEP will only have application to some system insofar as we can meaningfully define a generative model, giving the surprisal values that this system's trajectory is minimizing an action with respect to.

In the "state-based" formalism of the FEP, it was the steady-state requirement that both supplied a principled derivation of the generative model and a guarantee of its ongoing validity for that system. We derive the surprisal of a state in terms of how frequently the system has occupied this state in the past, and this is guaranteed to remain a valid measure insofar as a steady-state system is defined such that these frequencies cannot change. So, if the path-based formulation does not require any steady-state restrictions, as Ramstead et al. (2023) and Friston et al. (2023) claim, then how do we assign surprisal values to the possible states for an organism in order to assess if it does, or does not, comply with a least action principle?

One option is to wait until the organism has finished moving through the various possible states it might be in—that is to say, once it has died. Alternatively, we might be able to derive an equation of motion that describes the trajectory it will take by other means. Either way, once we already know the exact path a system took, or will take, then we can retroactively construct a probabilistic model, relative to which this path would qualify as minimizing the integrated free energy (Constant, 2021). This understanding of the FEP as a means of redescribing *any* trajectory in terms of Bayesian inference is reflected in Friston's occasional reference to the connections between the FEP and the complete class theorem, which states that there will always be some set of priors under which we can describe *any* behavior as approximately Bayesian (Friston et al. 2016).

The simplest way of redescribing any known trajectory as minimizing a free energy integral would be to incorporate time indexing into the states of the system. We can then note the trivial truth that, for each point in time, the system was certain to be in exactly and only the state that it was in fact in. We can then assign a negative log probability (i.e., surprisal) of 0, representing that certainty, to each of these states. If you already know the exact path I took through my possibility space on Friday, November 4, 2022, then, relative to this model there is 0 surprisal for the state of “writing about the free energy principle at 3 P.M.” and there is 0 surprisal for the state of “banging my head against my desk at 3:15 PM.” And look, just like that, you can now say that I took the path that minimized surprisal—insofar as I ever entered into the lowest possible 0 surprisal states. You can do all sorts of silly things if you redescribe certainties as probabilities—but it’s not a very productive use of time.

There may be more interesting means to retroactively derive a generative model under which a particular path would qualify as minimizing the integrated free energy over its entire course. As explored in active inference models of action planning, redescribing a known target trajectory in probabilistic terms may help with specifying a formal procedure by which a system could act to follow this trajectory, without itself knowing the full trajectory in advance. Insofar as this procedure could work for the retroactive description of *any* possible trajectory, however, it is merely a redescription and one that places no constraints on what a system must be like to be so describable. In contrast, the steady-state version of the FEP does at least propose some constraints on what sort of systems would qualify as free energy minimizers, and what they must do to continue to preserve their existence. It just turns out that those constraints don’t apply for exactly the living systems that its proponents target.

It is vacuous to claim that every trajectory will minimize surprisal if surprisal has no independent meaning beyond its assignment to make such a claim true. As described in section 4.1, the FEP has sometimes been presented in such “circular” terms, but this would not qualify it as an interesting extension of least action principles. Least action principles are also argued to be unfalsifiable, insofar as we may arbitrarily construct arcane alternatives to the standard Lagrangian of “kinetic—potential energy,” using standard physical quantities, such that we can describe any path as minimizing some function. Still, what makes least action principles interesting is how they relate physical quantities that have meaning outside of the principle. The difference between high kinetic energy may break your bones, but probabilities can’t hurt you. If either a path-based or state-based free energy principle is to say anything interesting about systems, living or otherwise, if it is to be of any use in modeling and predicting their behavior, then surprisal must be defined such that it can be independently quantified, prior to the assessment of whether it is subsequently minimized.

The free energy literature evolves rapidly, and book publishing is slow. By the time this book is out, even newer developments and revisions will certainly have emerged, which must remain lamentably unaddressed here. Still, whatever iterations and variations the free energy principle has undergone, or will continue to undergo, if these are to qualify as versions of the *same* principle, they must, at the very least, be defined in terms of probabilities. The only meaningful and non-subjective context we've seen for deriving such probabilities is in terms of steady-state systems, where we can identify some set of states whose frequency of occurring does not change over time. If the free energy principle is to be entirely freed from any steady-state requirement, as Friston et al. (2023) and Ramstead et al. (2023) suggest, then we would need an alternative basis for assigning these probabilities that is more meaningful than the post hoc redescription of an already known trajectory. In the absence of such a basis, we must continue to seek a steady state for at least some scale of biological organization, even if we no longer require it to apply at all scales.

9.4 Life Outside the Machine

So, if the free energy principle is to say anything meaningful about the systems that it purports to describe, then even the path-based formulation does not supply a means to get rid of the requirement for some sort of probabilistic stationarity. What it may allow us to do is to describe individual, *non-steady-state* systems as minimizing free energy with respect to a model that has been derived from a steady-state system at a different scale. This essentially provides a means of getting the multiscale approach back on the table, for we no longer require that every level of biological organization converge to a steady state. Instead, we might treat the individual in terms of whether its path-through states minimizes free energy relative to the prespecified "steady-state model" of its species.

This will only be meaningful, however, if we can derive a generative model at *some* level of biological organization—in order to fix our surprisal values prior to assessing whether some particular individual trajectory does, or does not, minimize them. As I will argue, we are no more likely to find this sought for stability at the level of the species or at any level of any organization that is composed of living systems. This is not *just* because biological systems are not, at any level, reducible steady-state systems, however. It is because they are not, at any level, machine-substances. This means that we cannot predetermine a fixed phase space and equations of motion for the process of living at either at the level of the species of the individual—no matter how multistable our attractor landscape, or how multilevel the equations in our non-trivial machine model (Rosen, 1991; Mirazo et al., 2004; Longo et al., 2012; Longo & Montévil, 2013; Koutroufinis, 2017; Kaufmann, 2000, 2019).

This would not just be a problem for the free energy principle, but for any attempt to capture living systems like other physical systems, in the standard formalisms of dynamical systems. If this is not possible, however, then it will undermine even the weakest and most trivial version of the FEP, insofar as this was committed only to the general idea that we can define *some* trajectory for a particular system, in order to subsequently redescribe it in surprisal-minimizing terms.

9.4.1 Instability at All Scales

In section 9.2, we considered one objection to the idea of modeling an individual organism in terms of a "species-level" ensemble distribution: namely that this would be too vast to be informative for the individual's trajectory. Even more troublesome is the fact that, despite having several with hundreds of thousands of years of evolutionary history under our belts, the species-level ensemble for humanity doesn't appear to have plans to converge into a stable pattern of behavior anytime soon. The possibility space of a species may change slowly, but it is still changing—for humans as much as any other species. Indeed, thanks to the rapid timescales at which we are continuing to transform our behavioral phase space through cultural and technological evolution, the ensemble distribution for humans is transforming within the life span of an individual person—not just across generations. For my great-grandfather, the possibility space for humans would have been very different from his first birthday to his ninetieth. Prior to the invention of the aeroplane, there was no chance of finding him, or anyone else, 33,000 feet in the air above the Atlantic Ocean. To be in such a position would be both nonviable and highly surprising by the conventions of the time. By 1990, however, transatlantic flight had become a much less improbable state.

Treating “altitude” as one of the variables making up the extended phenotype of human being is not ruled out by the FEP, which, as I’ve argued, provides no means to distinguish between essential and inessential states other than in terms of their stability. Indeed, we’ve seen that, just as Ashby did, advocates of the FEP have often explicitly presented the principle of generalized homeostasis as applying to all aspects of human behavior—as in Friston et al. (2020). Still, if we wish to give the example more physiological plausibility, we can reframe the example in terms of arterial oxygen saturation. In the early 1900s, a SPO2 of below 94 percent would have been an exceedingly uncommon occurrence. Since the late 1900s, however, it is becoming an increasingly common state for the many humans now regularly found cruising at an altitude above 33,000 feet (Humphreys, 2005).

A similar link between rapid cultural changes and average physiological states might also be made for blood pressure, with changes in average blood pressure of populations being consistently linked to increased industrialization—to the extent that the high prevalence of hypertension is sometimes referred to as a “disease of modernity” (Eyer, 1975; Dressler, 1999). Were we to adopt the free energy principle’s homeostatic equivocation of adaptive and average, we’d be compelled to describe a systolic blood pressure of 120–130 mmHG as adaptive for the specific class of the industrialized human—for all that every increase above 115 mmHG is known to lead to a range of increasingly nonbeneficial consequences, from fainting to heart attacks, stroke, or kidney damage (Lawes et al., 2004; Zhou et al. 2017).

Returning to my great-grandfather’s attempt to plan his life in terms of a generative model, we now see that he faced greater hurdles than his ignorance of variational approximation techniques. The problem was that at no point during his life had the human ensemble converged to a stable model that could be used to define probabilistic phenotype and that could be interpreted as minimizing surprisal with respect to.

This instability is not a unique problem resulting from some distinctive human capacity for information sharing. Nor can it be resolved by attempting a gene-centric “essential variables” response. Genetic variation does not just occur between organisms, via the copying of DNA for transmission from a parent to its offspring, it can also occur within the life cycle of an individual organism. Bacteria, for instance, trade genetic material like ’90s kids traded Pokémon cards in the form of small rings of self-replicating DNA called plasmids (Firth et al., 2018). Sometimes these plasmids can become integrated into a bacteria's central genome, and this horizontal gene transfer is one hypothesized origin for one of the three genes making up the *E. coli* Lac Operon (Hediger et al., 1985). Another possibility is the non-destructive insertion of genetic material by a retrovirus, the same mechanism responsible as much as 5–8 percent of the human genome (Belshaw et al., 2004), and potentially the development of the mammalian placenta (Mi et al., 2000).

Prior to the acquisition of this Lac Operon gene, there was no meaningful sense or scale under which either the lucky bacterium in question, or *E. coli* in general, had a high probability of lactose metabolism. There is no meaningful sense in which this would have been part of its phase space in any dynamical systems model of it. The acquisition of the ability to metabolize lactose is "surprising" on whatever timescale we attempt to locate a stable reference frame for this bacterium. The only way the FEP could describe the development and maintenance of this capacity is as a violation of the principle of free energy minimization, resulting in the destruction of one "generative model" and its replacement with a new one. Yet, rather than equating such a novel adaption to death, we would tend to recognize this as a development occurring within the same *E. coli* bacterium.

Now suppose that our hypothetical bacterium had never bumped into the hypothetical retrovirus that provided it with the final piece of the Lac Operon puzzle. Then *E. coli* might never have developed the ability to metabolize lactose that gives it the headstart on mammalian gut colonization, it might never have taken on its present role of preparing the ground for future waves of microbial immigration and we might not have developed the rich and diverse microbiota that is now responsible for synthesising a variety of vitamins essential to mammalian functioning, from the multi-purpose folic acid to the vitamin K required for the formation of blood clots (Maynard & Weinkove, 2020).

There are no necessary laws that entail the development of a microbiome. Some animals lack one altogether (Hammer et al., 2019), and there are mammals, such as bats and pandas, that do not appear to depend on microbial support to meet their nutritional needs. Still, most mammals do, and when specimens, known as "germ-free" lines, are produced without this microbial colony, studies show that their nutritional demands in the absence of this assistance are up to 30 percent higher (Wostmann et al., 1983).

Translating this increased nutritional dependence to non-microbially assisted early hominids, it is not implausible to suggest that the need for higher levels of leafy vegetables to meet our vitamin K requirements might have constrained our adaptability to less fertile regions, and limited or slowed their outward expansion. And who knows what diverse and terrible beasties might have evolved across the Americas within the extended grace period afforded by the delayed arrival of the first ecosystem-destroying Homo sapiens across the Bering Stait (Smith et al., 2018).

All this, for want of one missed connection between an *E. coli* bacterium and a retrovirus.

Like most attempts at narrating deep evolutionary history, this tale of sliding genetic material is something of a just-so story. Stuart Kauffman (2019) provides a similar account in his book *A World beyond Physics* (p. 97) describing the evolution of the first food chain, and the first symbiotic relationship among protocells. The endosymbiotic origin of the mitochondria that supply eukaryotic cells with the improved energy efficiency that powers their increased complexity is perhaps the most famous such lucky accident. Another more complicated narrative might be given for the retrovirally assisted development of the mammalian placenta and the transition from oviparity (egg-laying) to viviparity (internal gestation). Perhaps such events would have inevitably occurred *eventually*, but the point is that *when* they actually did happen, and the particular context in which they occurred, determined how such events either facilitated or foreclosed various further developments with consequences far beyond their immediate effects on an individual organism.

Such stories may inevitably be speculative, but hard as it is to trace back the sequence of events that led to where we are now, it was much harder to predict them in advance. Indeed, as Longo et al. (2012) claim, the inability to locate some rule, or symmetry, that must be preserved through any transformation is what makes organisms "unprestatable"—that is to say, unformalizable with the tools of dynamical systems, or indeed any other machine-type logic. As they argue,

> Thus, it is proliferation, variation and selection grafting novel phenotypes into evolving organisms that reveals, again after the fact, the newly relevant and unprestatable observables and parameters. Thereby, this is our main thesis, the very phase space of evolution changes in unprestatable ways. In consequence, again, we can write no equations of motion for the evolving biosphere, nor know ahead of time the niche boundary conditions so cannot integrate the equations of motion which we do not have. No law entails the evolution of the biosphere. (p. 6)

Longo et al. focus on how these unprestatable developments unfold over phylogenetic timescales—and thus prevent the formulation of exceptionless laws describing the evolution of some population, species, or indeed the biosphere as a whole. More important here, however, is the fact that these symmetry-breaking events do not only occur *between* generations, from copying errors and recombinations in the transmission of DNA from a parent to its offspring, but also within a single individual. Events such as horizontal gene transfer or rapid within-generation changes to an individual's cultural and physical niche may radically alter the trajectory and state space for that organism *within* its lifetime.

By way of explaining how the machine concept of the organism fails here, it is not enough just to say that an organism is a process, not a substance. This does not itself explain what it is that differentiates organisms from other, less troublesome sorts of activity like countdowns or rotations, for which we can still provide a fixed rule that describes their unfolding. As Longo and Montévil (2014) point out, "The dynamics of biological organisms, in their various levels of organization, are not 'just' processes, but permanent (extended, in our terminology) critical transitions and, thus, symmetry changes" (2014, p. 161). The key difference between physical and biological processes, as they argue, is that "usual physical processes preserve invariants, whereas extended critical transitions [characteristic of biology] are a permanent reconstruction of organization and symmetries, i.e., of invariants" (p. 175).

To illustrate this point, consider again a pendulum. Here, we have a set of variables, such as angular displacement and velocity; parameters, such as string length and pendulum mass; and background constants such as the gravitational constant. We also have some symmetry principles and conservation laws that specify things that must remain invariant as the system moves through a fixed phase space of possible states. For instance, conservation of energy tells us that when potential energy increases as the pendulum's angular displacement increases, so angular momentum must decrease. This allows us to derive a set of equations to describe every state this system might possibly enter while still remaining an (undriven, frictionless) pendulum (see figure 9.1).[2]

A real pendulum is not an isolated system, and processes not included in our model might disrupt it in ways that our model cannot predict. But while we cannot predict these from the study of the pendulum alone, we can still specify every possible position in phase space that it might unexpectedly

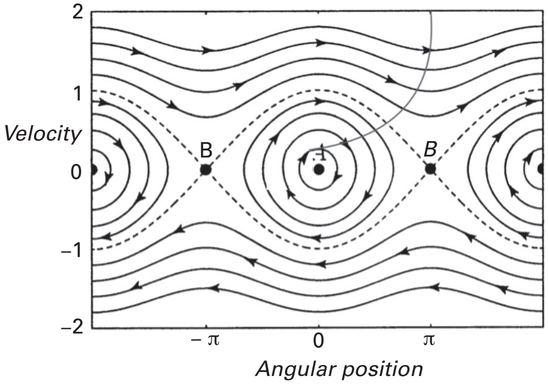


Figure 9.1
Phase space for a simple, frictionless, pendulum illustrating an "impermissible" flow in light gray (adapted from Winter & Murray, 1997).

be displaced into, along with how it will behave subsequently—*so long as it remains the same type of system*. If we allow our parameters to be varied, we can also fix what would happen if the string were extended or if we increased the weight of the bob.

This doesn't mean that the pendulum in itself contains all the information we need to exhaustively specify every possible thing that might happen to its constituent parts. From studying our pendulum alone, we could not prespecify what would happen if the experimental laser in the much more exciting laboratory next door were to malfunction, sending a beam of super-high powered coherent light through the wall and vaporizing our modest little experiment into disconnected gas particles with greatly expanded degrees of freedom. Still, what we can do is give an exhaustive specification of the possible trajectories that are compatible with that pendulum continuing to exist *as a pendulum*.

For the pendulum then, the possible ways it might be affected by the world are exhausted by either perturbation of its position in a fixed phase space or its destruction. In contrast, because an organism is not individuated by a set of fixed rules but rather by some relation between its temporal stages, so an externally induced event—whether the insertion of retroviral genes, a plasmid exchange, or the learning of a new skill—may alter the equations governing its dynamics and expand its phase space in a way that we do not necessarily consider as the destruction of that organism.

Is this true unpredictability? Is it due to something about the metaphysical status of organisms rather than just a reflection of our own epistemic limitations? Longo et al. (2012) propose such an indeterminist view, taking the unprestatability they ascribe to biological systems to be, at least partly, grounded in genuinely random events, pointing to the possibility for indeterministic quantum mechanical events to trigger point mutations that could have phenotypic consequences for the organism. Yet, the fact that the macroscopic trajectory of an organism might be altered by metaphysically unpredictable events cannot be the only thing that makes organisms "unprestatable" in a way that machines are not. Thanks to the complexity and small scale of modern processing components, many computational machines are vulnerable to what are known as "single event effects"—triggered when a single randomly released particle, such as a photon, crashes into a single crucial transistor and changes its state—with macroscopic consequences ranging from the crashing of aeroplanes to the overturning of elections.

While it is important for my purposes that the unprestatability I'm ascribing to the trajectory of an individual organism reflects a genuine difference between that organism and a machine, not just a matter of our epistemic limitation in light of the greater complexity of the organism, I don't think the genuine randomness of quantum mechanical upsets is the only way to draw this distinction. What matters most in this respect is not just the randomness of *when* an event happens, but the relational nature of *what* it means for the trajectory of that organism. A functional safety engineer, tasked with designing integrated circuits for critical purposes, has no way to predict if, or when, a single photon will strike a particular transistor in that circuit board—but their job is contingent on the fact that the fixed structure of that circuit allows them to specify exactly *what* will happen if it does. What distinguishes an organism from a machine is not the possibility of its being macroscopically disrupted by subatomic fluctuations, but rather the way in which environmental changes not only destroy or perturb organisms but also *enable* them by expanding their repertoire of possible behaviors. Whether macroscopic changes in an organism's structure are deterministic in origin or not, the point is that their consequences for that organism depend on how they interact with a host of other changes within the rest of the organism, in its environment, or with other organisms.

We can make a generic statement about the consequences of lengthening the string in a pendulum without reference to any particular pendulum or its environment. Rolling back the clock billions of years, the same would not be the case for the consequence of the insertion of a lacA gene into an *E. coli.* Whether that insertion would lead to the capacity for lactose metabolism will have depended on *this particular E. coli* being one that had also evolved the other lac operon genes, and whether it would lead to faster mammalian gut colonization depended on whether there were milk-guzzling mammals waiting around with guts ready and waiting to be colonized. Similarly, the occurrence of genetic changes resulting in the development of bony fins of an ancient aquatic tetrapod would not, in themselves, have allowed us to predict that the tetrapod would walk on land. This depended also on the drying out of the environment in which that tetrapod lived, something that could not have been predicted solely by studying the structure of that tetrapod alone. Bony fins played a causal role in the development of amphibiousness, but they did not entail it. Prior to the arrival of dry land into its environment, walking around was not a part of the possibility space for that tetrapod.

So, making sense of the possibility of transformation, in addition to perturbation or destruction, depends on a rejection of the machine-substance conception of an organism as being individuated by some invariant structure and dynamics. For these sorts of reasons, Longo et al. (2012) argue for the enablement of unspecified possibilities, as more apt than the entailment of some specific effect, for the study of biological causality. This move alone is not enough to secure a claim of objective unpredictability for the development of a biological system. Without genuine randomness in the generation of macroscopic mutations, a Laplacian demon could still predict the development of an individual organism, in virtue of its ability to trace every interaction that occurs in the world as a whole. But this possibility of enablement of new behaviors is enough to distinguish the organism from the pendulum insofar as it prevents even this demon from prestating a finite set of possibilities for how the former could develop (conditional on its not being destroyed) in isolation from knowledge of the development of the entire environment in which it is embedded.

The argument that we cannot pre-state the phase space and equations of motion for an organism may sound like biologists arguing against the very possibility of doing biology. But it should not be taken as a blanket objection to the use of models, such as those of dynamical systems theory, to describe organism dynamics. We can still describe stereotypical behaviors of an organism using dynamical systems models, and we can retrospectively construct a model of the trajectory that an organism did in fact take. To take a processual view just means remembering that these dynamical models are only locally valid approximations, subject to change in ways that we cannot predict in advance. In this context, the bold claims about "existential imperatives" made by advocates of the FEP serve as a nice *reductio ad absurdum* against mistaking the relationship between some particular, contingent pattern of behavior and a mathematical model of it, for an identity between the behaving organism and said model.

Moreover, to say the biological constitutes an unprestatable *World Beyond Physics* as Kauffman does is not to say it is a supernatural world beyond the physical, populated by vitalistic entities and energies. It is just to say that it is a realm that cannot be predicted with the sorts of physical models that are currently dominant. This should not be controversial. The various inadequacies of the Newtonian paradigm of a mechanical universe are well established and, as Rosen (1999) notes, complaints about the impoverished tools of the field have been made by leading physicists throughout history, from Schrödinger to Einstein. But as each new mathematical formalism improves our ability to describe some aspect of a system's behavior within limited and arbitrary constraints our enthusiasm can lead us to forget all the phenomena beyond those constraints for which it fails.

So, both the free energy principle and a machine substance ontology more broadly, fail to capture the ways in which organisms are free of allegiance to any particular material instantiation or governing equations. Before moving to whether enactive accounts can do any better, I want to briefly consider whether a more modest form of the FEP, uncommitted to machine-substantialism might be useful instead. Once we separate it from this machine-substantialist framework, I believe we can see why a description of living systems cannot afford to ignore the metabolic nature of organisms.

9.5 Life as the Process of Seeking Stabilities

Rather than taking the FEP as providing a definition of a living system in terms of a fixed generative model, we might instead opt for the weaker claim that the life cycle of a living system is characterized by the need to continuously reestablish *some* steady states, if not the mandatory preservation of the *same* steady state. This would be to give up the guarantee of a particular joint probability density describing the unchanging phenotype by which a particular organism could be identified. All it would allow us to say about the overall trajectory of an organism is that it will move between states or patterns of behavior that it can stably maintain. This releases free energy minimization from an essentialist commitment to invariant features, while still allowing it to describe an imperative that a system must satisfy.

The problem with this approach is that, under the FEP, all steady states are treated equally. The principle itself makes no discrimination between the viability of possible steady states, other than in terms of how stable they actually end up being. A stability-based analysis of survival applies well to the process of death and consequent decomposition—which are well described in terms of a failure to secure stability—but this doesn't mean that any stable state will automatically be good for the organism in virtue of that stability alone.

An alternative response to declining food stores in some species of bacteria is to enter a frozen state of cryptobiosis, called an endospore, in which they can persist for thousands of years. Similar strategies are encountered in other microorganisms, like the famously resilient tardigrade (Wright, 2001) and even animals as large as Alaskan wood frogs, which can spend up to seven months frozen solid (Costanzo et al., 2013). In extremely rare cases, this kind of self-mummification may even be practiced by humans, as with the practice of Sokushinbutsu among Shingon Buddhist monks in Japan or cryogenesis among Silicon Valley billionaires—though a rather crucial difference is that these transitions have never been successfully reversed.

Whether or not cryptobiotic organisms, whose vital processes will restart under more auspicious conditions, should be counted among the living while in their frozen state is not so much controversial, as it is oddly undiscussed by philosophers of biology. A scattering of remarks by biologists studying the phenomenon suggests that many hold that "on an organismal level, they are essentially dead," as Alaskan wood frog researcher, Don Larson puts it (quoted in Netburn, 2014). Such assessments, as Neuman (2006) argues, are typically based on the presumption of ongoing metabolic activity as a necessary criterion of life. That there is any question about the matter can be credited to the fact that in several cases, such as the tardigrade, it is difficult to establish whether metabolism is truly completely inactive (Pedersen et al., 2020). "Crypto" means hidden, not absent, and Keilin (1959) initially introduced the term to refer specifically to the absence of "*visible* signs of life" (Clegg, 2001, p. 213). For the above biologists at least, determining whether the cryptobiotic tardigrade is alive or not looks to come down to the empirical question of whether it is, in fact, continuing to metabolize residual energy stores at undetectable levels.

We can contrast this metabolic understanding of life with the FEP, which can make no distinction between stable states that depend on ongoing metabolism and those that do not. The proposal that organisms are driven to just establish *any* stable state fails as a normative principle for living systems, insofar as it treats both metabolically inactive phases and the kinds of steady state we normally associate with being alive as of equivalent value. The requirement to be in some steady state has nothing to say about why we (or at least I) would be inclined to say that the *E. coli* chowing down on lactose is doing rather better than the frozen endospore, why we'd be inclined to describe the thrill-seeking roller-coaster rider as "more alive" than the self-mummified monk—despite the impressive stability of the latter. Some transitions between steady states may not equate to the death of the organism, but others do. What is needed is a principled account of why some steady states count as viable and others do not, why some transitions are compatible with the continuation of an organism's existence, and others are not.

This is precisely what an account of the autonomy of an organism that does not abstract away from metabolic self-production, should provide. A state is viable for an organism only insofar as it is compatible with the particular self-producing organization of the organism at that time. So, the organization of a system at a time will produce limitations on which states are viable and which are not, and *for so long as the organism's organization does not change*, it will need to regulate its states in such a manner that they remain within these limits. This regulation may be homeostatic and may be described in terms of free energy minimization.

Crucially, however, this organization need not remain invariant for the organism to continue to live. What allows us to follow the organism through time is not that this organization has any fixed form, but the way that each transformation is related to the next in terms of the relationships of reciprocal dependence between the organization, the processes it enables, and the organization they produce in turn, and so on. As the particular form of this autonomous organization changes, so what states are viable for it may also change in ways that cannot be prestated by the presumption that it, like a machine, has an organization that remains invariant.

So, the best we can say of free energy minimization is that it is a way of approximating a contingent, if common, feature of biological dynamics. That's still pretty useful! We don't only want to describe the necessary constraints on biological survival; we also want to model what organisms often tend to do and how they do it. I see no reason why active inference accounts of free energy minimization couldn't be useful in this respect. Still, an account of autonomy that can adequately characterize what is necessary for survival must come first. In the next chapter, I want to take a closer look at bioenactive characterizations of autonomy, to see why the standard formulation is not quite up to this task, before suggesting an alternative that I believe is, and which I argue the bioenactivist should adopt if they want an adequate, non-trivializable grounding for talk of biological autonomy, teleology, and intentionality.

10 Biological Autonomy

In being uniquely unmoored from either material or organizational commitments, organisms are not machines and cannot be reduced to the logic of machines—for all that it may sometimes be useful for explanatory purposes to model them in mechanistic terms. The claim that an organism's capacity for change is not restricted in advance by any invariant dynamical equations, no matter how higher order they may be, marks a genuine distinction in kind, not merely degree of complexity. But the fact that we can follow an organism throughout these changes leaves a puzzle as to how we do so. Moreover, a distinction in terms of the capacity for open-ended change is not enough for bioenactivism, as it does not tell us what marks organisms alone as *agents*.

The answer to both of these questions, I believe, lies in choosing the intrinsic instability of a metabolically produced system over the default stability of a homeostatic one as our starting point for a theory of the organism. This priority given to the preservation of autonomous organization over the homeostatic preservation of stability is, Di Paolo et al. (2022) argue, a distinguishing feature of the enactive approach (or what I'm referring to more specifically as bioenactivism) which raises a barrier to any potential compatibility with the free energy framework. Still, the degree to which this preservation of autopoietic or autonomous organization is compatible with open-ended development has not always been clear.

Indeed, autonomy and autopoiesis are defined in Maturana and Varela's early work in explicitly machine-like terms, where the cell is presented as an "autopoietic machine." As Thompson (2007) emphasizes, this does not reflect the contemporary, atomistic concept of mechanism as something whose behavior can be decomposed in terms of local interactions between its parts, but rather describes a system with a relational organization that is multiply realizable and independent of its particular instantiation (what Maturana and Varela refer to as its structure) at a particular moment. Nonetheless, this commitment to an invariant organization still fits within the "essentialist" view of the machine-substance conception—where, rather than being defined by the first-order properties of its parts, an organism would be individuated by some invariant second-order relations between these that do not change over time. In other words, as Di Frisco (2014) notes, Maturana and Varela have not so much rejected substantialism but moved from an atomistic understanding of substance to a formalist one, allowing material instantiation to vary, while holding that it is instead the form (or "organization") that must remain invariant.[1]

This is the same move that was discussed in the previous chapter for allowing the FEP to account for the turnover of material components making up its fixed causal graph. As described, talk of higher-order organizational invariants can also allow us to make sense of lower-level variations in dynamics, as when an organism moves from one pattern of behavior to another. The suggestion that the equation that directly governs a system's state transitions is not fixed is much weaker than Longo et al.'s (2012) statement that "there are *no* entailing laws" or fixed equations of motion for biological systems, however. Even if we allow that the higher-order equation governing the changes in this lower-level state-to-state transition function can itself also change, even if we introduce arbitrarily many layers of higher-order transition rules to account for this, the "higher-order invariant" view is committed to the idea that this hierarchy of equations will ultimately terminate with some fixed rule that governs all the changes at the levels below.

It is this fixed rule that corresponds to the "organization" that, for Maturana and Varela, defines a living system throughout these "structural" variations in lower-level rules. But, insofar as we accept the arguments of the previous chapters for the unprestatability of an organism's phase space, then for the organism, unlike the non-trivial machine, there is no such termination, no fixed rule and no fixed organization that we can reduce all changes in structure or behavior to variations within.

The consequence of viewing organisms as machines, rather than unprestatable processes of critical transitions, is illustrated in Maturana and Varela's replacement of an entailing relationship between input and output with the cybernetic concept of a "perturbation"—a compensatable disruption to an otherwise stable mode of being. Contrast this with Longo et al.'s (2012) suggestion of enablement, as more apt than entailment for the analysis of biological causation. Unlike reduction of events to "perturbations," which reduces the environment to a source of disruptions and challenges that one must preserve one's inherently stable form against the idea of enablement presents our changing surroundings as an expanding well of possibilities for continuing a process of production and individuation in novel and unpredictable ways.

All of this is not a problem for the goals of autopoiesis theory. As discussed in chapter 7, Maturana and Varela were not initially aiming to distinguish organisms from machines. The residual idea of an invariant organization that is continually reproduced has, however, created concerns about any claimed compatibility between bioenactivism with processual accounts of the organism (Di Frisco, 2014), raising the question of whether, as Meincke (2019) asks, autopoiesis might be "a substance wolf in process sheep's clothing?"

While, as described, definitions of autopoiesis have often been framed in terms of the regeneration of some invariant set of components, processes, or relations, my categorization of bioenactivism is not wedded to the specifics of the early formulations. The importance of autopoiesis, from my perspective, is not that it provides a sacrosanct analysis of the necessary and sufficient conditions for life, but that it points toward a different approach to explaining what a living system is in terms of the logic of self-production—rather than the capacity for differential reproduction and evolution, or in terms of the particular and contingent chemical form of it that is familiar to us.

Moreover, as Di Paolo et al. (2022) argue, a distinguishing feature of the *enactive*, as opposed to merely autopoietic, concept of an organism is the focus on "precarious, self-constituted entities in ongoing historical development and capable of incorporating different sources of normativity, a world-involving process that is co-defined with their environment across multiple spatiotemporal scales and together with other agents" (p. 3). It is this focus on cumulative historical change, they argue, that constitutes one of the irreconcilable theoretical tensions that undermine claims of potential compatability between the steady-state formalism of the free energy framework and the enactive approach.

As an evidential basis for this account of the importance of historicity in the origins of the enactive approach, they point to Varela's statement in his preface to the 1994 reissue of his and Maturana's *On Machines and Living*

Beings, where he acknowledges some inadequacies in their account because, as he puts it, "it seems to leave the phenomenon of interaction in a grey area of being a 'mere' perturbation" one that, as he puts it, "does not properly take the account of the emerging regularities in the course of a history of interactions. Accordingly, he introduces an alternative to his and Maturana's concept of "structural coupling," with a view to, "turning historical reciprocity into the clue of a co-definition between an autonomous system and its environment. I propose to call this point of view in both biology and cognitive science, *enaction*" (2011/1994, p. 614).

So, as Varela recognizes and a number of authors have argued, these enactive ideas do not derive immediately from the original formulation of autopoiesis. A recent concern of enactivist literature has been on how to supplement, or modify, the concepts of autopoiesis and autonomy to render them more suitable for bioenactivist aims. A particularly significant development in this respect, as briefly mentioned in the introductory discussion of bioenactivism, is Di Paolo's (2005) claim of the need to supplement autopoiesis with an account of adaptivity, describing the organism's capacity to regulate its coupling with its environment. As mentioned in chapter 1, this idea of adaptivity, and its relationship to an open-ended process of learning, has since been developed in Di Paolo, Buhrmann, and Barandiaran's use of Piaget's ideas on sensorimotor equilibration to describe how novel interactions can be incorporated into agent's repertoire.

Another particularly interesting move toward a more processualist conception of the organism is the uptake of the work of the twentieth-century French philosopher Gilbert Simondon, among both critics of autopoiesis (Di Frisco, 2014) and enactivists (Di Paolo, 2020). In rejecting the attempt to define an individual in terms of either a fixed form or fixed material parts, Simondon advocates prioritizing the process of individuating within which each individual is but a temporary phase and which, for the organic, is an inherently unfinished process. In this respect, Simondon's work stands as an interesting philosophical precursor to the arguments of complexity theorists and theoretical biologists reviewed in this chapter, and, as Di Paolo argues, is well suited to supplementing the bioenactive literature by directing a focus toward this "open-ended becoming."

Still, in discussions of this kind of adaptivity and accommodation, the focus is on the sensorimotor level—suggesting the presumption of a more basic network of bodily metabolic processes that must be preserved and, in service of which, ever new sensorimotor engagements might be incorporated. This is clearly an improvement on the FEP's language of steady-state ESIA cycles, insofar as the bioenactivist at least has the means to describe a distinction between the preservation of a particular network of self-perpetuating processes, versus the open-endedness of the possible sensorimotor engagements that might serve to preserve it. Moreover, unlike Ashby's starting point in the homeostasis of "essential variables," the bioenactivist can give a principled account of *why* some particular variables must be kept within particular bounds, whereas others can be freely varied in support of this goal, in terms of how the former must be kept within those bounds if the autonomous network of processes is to continue its operation.

Nonetheless, as a consequence of a focus on these secondary dimensions of autonomy in recent enactive work, with the aim to scale up to explanations of cognitive processes of learning and development, I believe descriptions regarding the prior autonomy of the organic body—the constitution of which autopoiesis was intended to pick out at the molecular level—remain insufficiently refined.

As described in chapter 1, the definition of autonomy that I take to be the widely accepted one in contemporary bioenactivism is given in terms of an operationally closed network of processes, such that every process both depends on at least one other process in that network and enables a further process, together with the requirement of precariousness, such that these processes would not continue to operate outside of said network. As De Jaegher and Di Paolo state this,

> An autonomous system is defined as a system composed of several processes that actively generate and sustain an identity under precarious conditions. To generate an identity in this context is to possess the property of operational closure. This is the property that among the enabling conditions for any constituent process in the system one will always find one or more other processes in the system (i.e., there are no processes that are not conditioned by other processes in the network which does not mean, of course, that conditions external to the system cannot be necessary as well for such processes to exist). By precarious we mean the fact that in the absence of the organization of the system as a network of processes, under otherwise equal physical conditions, isolated component processes would tend to run down or extinguish. (De Jaegher & Di Paolo; 2007, p. 487)

Equivalent formulations are also given in Di Paolo (2009), Thompson and Di Paolo (2014), and Di Paolo et al. (2017). To avoid debate over the difference between early characterizations of "operational closure" (Bourgine &

Varela, 1992) versus what Bich and Arnellos (2012) argue should instead be referred to as "organizational closure," I will refer to this characterization as "*process closure*." Like all accounts of closure, this does not mean that the system with process closure does not depend on other processes external to it—only that from amid all of these dependence relationships we can extract a network of *mutual* dependence. And it is only those processes that both enable and are enabled by other processes within this network that will be a part of the system that realizes process closure.

As depicted in figure 10.1, this definition of process closure tells us how to extract those processes that are, or are not, a part of a particular set of cycles that occurs over some chunk of time (see figure 10.1). Importantly, however, the recurrence time of many of these cycles is shorter than the life span of an organism. The regeneration cycle for liver tissue, for instance, can be as short as a few days (Sender & Milo, 2021). While there are also much longer cycles making up the organism, as argued in the previous section, we cannot conceptualize the identity through time of the organism in

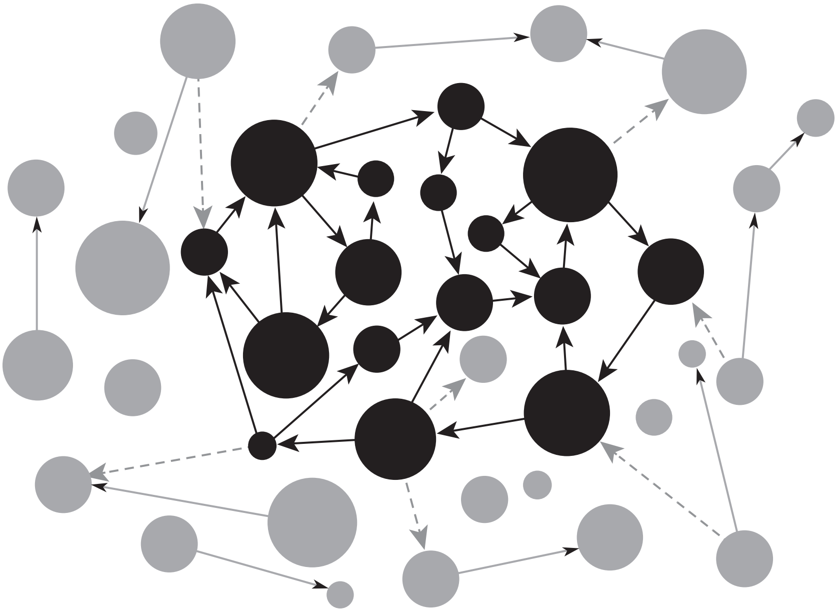


Figure 10.1
An illustration of how an operationally closed network of processes (highlighted in black) is distinguished from its surroundings (Di Paolo, 2013).

terms of one longer overarching cycle. A more obvious problem with doing so is that every organism's life span is terminated in the inevitable failure to "loop back" when the organism dies.

So, process-closure only tells us the spatial extent of an organism over the particular period within which a set of cycles occurs. The question is how we identify the organism throughout multiple cycles. Must its identity be defined by the *same* repeating cycles, the same organization, the same processes, and the same products? Di Paolo et al. (2017), for instance, often talk of a cycle of regeneration among the same set of processes in the same network, implying that beneath the open-ended adaptivity of sensorimotor learning we might still be able to identify a stable, essential organic core by which the "self" of the individual organism might be defined. Given the discussion in the preceding sections of this chapter, and, specifically, the metabolic plasticity observed not only in single-celled bacteria like the *E. coli*, but even in multicellular organisms like deep-sea fish (Raposo de Magalhães et al., 2021), I am not convinced that it is possible to specify an invariant network of processes at a level both specific enough to individuate a particular organism and flexible enough to incorporate the possibility of such changes.

Nonetheless, the above definition of autonomy seems compatible in principle with a view in which the organism's identity through time is determined not by the preservation of any one particular network of processes or components, but in terms of the continuity of relations of production between a series of evolving process-cycles. Though Austin (2020) is critical of the attempt to avoid resorting to substantialist principles in characterizing an organism, he notes, "Amending the rather permissive relation of genidentity with the *constructive* character of autopoiesis certainly furnishes one with a more restrictive criterion of slice-series composition" (p. 9). The same might be intended with autonomy, such that even if a particular process network underwent organizational change, we would still be able to take the new organization and the old one as manifestations of the same organism, thanks to the latter organization being dependent on the operations of the former for its existence.

So, the fact that the enactive approach has been concerned with developmental change since its beginnings is enough, as Di Paolo et al. (2022) argue, to undermine claims of compatibility with the FEP. I believe we can make a stronger claim than incompatibility, however. I have argued that this non-cyclical behavior is a pervasive and distinctive feature of living systems that the FEP's language of steady-state ESIA cycles is constitutively incapable of describing. This motivates a ruling against its ability to serve as a "first principle" for living systems—enactivist or otherwise.

Still, the bioenactive definition of autonomy is not without its problems. If we are to individuate the organism in terms of relations of production between temporal stages, rather than in terms of an invariant organization that is reproduced, then we need a robust account of what these relations of production are. As I will argue in the next section, process-closure is inadequate to capture the relations that underpin a self-producing continuity and, as such, is inadequate to ground the attribution of purposiveness and intentionality to living systems.

10.1 The Problem with Defining Autonomy as Closure of Processes

The second aspect in which I argued that the FEP failed as a theory of living systems was in how it downgraded material turnover to optional changes that *may* occur within the constraints of a necessarily invariant organization. As I have argued, it is only by describing how the organism is not merely free of a particular material basis but in *need* of a continual flow of matter by which it continuously reconstitutes itself, that we can adequately characterize the precarious dependence of organisms on their own activity. It is this characterization that we need to properly ground the bioenactive understanding of agency, intentionality, and immanent teleology (Jonas, 1953, 2001/1966; Weber & Varela, 2002; Thompson, 2007).

So, one improvement of the contemporary bioenactive definition of autonomy over Bourgine and Varela's (1992) operational closure is that it describes an ongoing process of individuation under which change and process are primary. Nonetheless, as Bickhard (2008) argues, the material and energetic conditions of life remain underdeveloped in these contemporary formulations of autonomy. This leads to the risk of trivialization, for, as Thomson and Di Paolo (2014) acknowledge,

> all material processes are precarious if we wait long enough. In the current context, however, what we mean by "precariousness" is the following condition: In the absence of the enabling relations established by the operationally closed network, a process belonging to the network will stop or run down. (p. 4)

This clarification is not sufficient for precarious operational closure to pick out living systems distinctively. As Moreno and Mossio (2015) and

Mossio and Bich (2017) describe, many physical and chemical systems from tornados and convection rolls to the hydrological cycle would meet this set of requirements.

Consider a simple network of a swing ball set and a robot programmed to play it. Here we have two mutually dependent processes. The first: the balls orbiting around its pivot. The second: the robot's moving its arm to hit it. These processes do not merely enable each other, they depend on each other. If the robot doesn't hit the ball, the latter's rotation would run-down, to leave it hanging limply from its string. If the swingball didn't continue to swing around, the robot's hit-the-ball process wouldn't be activated because there would be no ball swinging past that it is able to hit.

Here we have a precarious and operationally closed network of processes. If this were all there is to autonomy, then, with additional capacities, such as the ability of the robot to regulate the force of its stroke with respect to wind conditions, we would have a system that is also adaptive and thus, something that starts to look like it meets Di Paolo et al.'s (2017) criteria for being an agent. I don't think we should accept this. To take the coupled robot-swingball or the hydrological cycle as autonomous, as intentionally oriented toward goals and norms, is to invite the criticism, oft raised against bioenactive approaches from traditional autopoiesis theorists (Villalobos, 2013; Villalobos & Ward, 2015; Villalobos & Dewhurst, 2018), that we are inappropriately projecting teleology on to systems that are merely state-determined mechanisms. Moreover, I take this insufficiently robust account of self-production and precariousness to be what has left bioenactive work vulnerable to the kind of trivializations of autonomy presented by the FEP, under which even coupled pendulums and Watt governors are argued to qualify. Without a robust notion of precarious self-*production*, we cannot distinguish the adaptive agency we might wish to ascribe to organisms from the capacity of a coupled feedback mechanism to change its state or structure, as part of a continual cycle of activity.

This doesn't mean we should abandon the bioenactivist program to opt for either instrumentalism or hylozoism. All it means is that bioenactivists have not sufficiently captured what is special about the metabolic self-production of a cell in virtue of which it carries the germ of intentionality, teleology, normativity, and agency.

10.2 Constraint Closure

In what respect has the above definition failed to capture the thermodynamic basis of autonomy, as Moreno and Mossio (2015) suggest? The cyclical process of a robot-swingball set depends on a continual flow of energy into the system from outside, via the robot's power cord, but mere energy dependence is trivial. All processes depend on a flow of energy for their continuation—the miracle would be a system where they did not.

There are two key respects in which the robot-swingball is nothing like an organism. The first is that neither the robot-hitting nor the swingball-swinging play any part in securing or regulating the energy supply that enables them. The second is that the *structure* of both the robot and the swingball set are intrinsically stable and will persist with or without this supply. It is only their dynamics that depend on a flow of energy (as indeed all processes do); the structure of the robot-swingball set does not.

The problem, as Moreno and Mossio put it then, is that this account of closure among processes "fails to locate closure at the relevant level of causation" (Moreno & Mossio, 2015). Like the analogy of food as fuel discussed in section 8.2, it presumes the system can be factored into a fixed set of background constraints and the "autonomous" dynamics that result. These fixed constraints are taken for granted, and their presence is presumed to require no explanation when we come to describe the system any more than information about the manufacturer is required for a dynamical model of a pendulum. It is this presumed division, as Koutroufinis (2017) describes, that is the key move in dynamical systems theory, where we separate the invariant equations of motion, determined by the constraints of a fixed organization, from the resultant dynamics, which cannot alter those constraints in turn.

This modeling division is not a distortion when what we are modeling is a machine. We really can "divide" a machine's structure from its activity by stopping the machine, and we can also pull those structural pieces apart. We can turn our car off and start it again and we can take it apart and put it back together again without consequences. This cannot be done for organisms because these constraints are themselves dependent on the processes that they collectively enable for their continuance (Dupré & Nicholson, 2018; Nicholson, 2018; Mossio & Bich, 2017). As such, Montévil and Mossio (2015) propose that the appropriate characterization of the organism is in terms not of closure of processes but more specifically closure of *constraints*—an idea inspired by Kauffman's (2000) work on autocatalytic sets and work-task cycles on the one hand and by Rosen's (1991) ideas of closure to efficient causation on the other.

Their account is elaborated at greater length in Mossio and Moreno's (2015) book *Biological Autonomy*, where it is defined as follows:

> In formal terms, a set of constraints *C* realizes closure if, for each constraint Ci belonging to *C*:
>
> 1. Ci depends directly on at least one other constraint of *C* (Ci is dependent);
> 2. There is at least one other constraint Cj belonging to C which depends on Ci (Ci is enabling). (p. 20)

Before explaining whether this account is sufficiently specific to the organic and whether it has the potential to provide a naturalistic grounding for teleological, intentional, or normative talk, we need first to say a little more about this distinction between constraints and processes, which Mossio and Moreno define as follows:

> Processes refer to the whole set of physicochemical changes (including reactions) occurring in biological systems, which involve the alteration, consumption and/or production of relevant entities. Constraints, in turn, refer to entities that, while acting upon these processes, can be said to remain unaffected by them, at least under certain conditions or from a certain point of view. (p. 11)

This point is cashed out more formally in terms of symmetries, where a constraint is something that remains invariant with respect to particular thermodynamic flow, just as the length of a (frictionless) pendulum string, or its overall energy, remains invariant throughout its changes in angular position. The first natural question to arise here is how can something that is partially defined by *not* changing also be something that "acts upon" or causes something else? The second concern is how something that is defined in terms of its invariance could, at the same time, be precarious and in need of regeneration? For a constraint to be a cause, or to be an effect of some process, is actually quite common outside the biological realm. I will deal with each in turn before describing how they can connect together in the structure of constraint closure unique way to biological systems.

10.2.1 Constraint Causation

The standard philosophical view of causal relata, as Schaffer (2016) describes, is that they are events. One change occurs and, in a law-governed way, triggers a subsequent change as its effect. Additional factors, like the fragility of glass or the energy contained in sugar, may alter what happens, but they do not tell you why thing happened at a particular time. So, when your partner asks, "Why is my favorite glass broken?" they are asking what caused the event of its breaking and are expecting to be answered with some event—perhaps involving your recent tendency to conduct chemistry experiments with wanton disregard for the integrity of other people's kitchenware. Pointing to "thermodynamics" is seldom an acceptable response.

This is the view of causation offered in the mechanical conception of the universe, inherited from Newton, as essentially a collection of inert and independent billiard balls, that change their velocity only when bumped into by another in a manner determined by some external and eternal laws. But this is not the only way to think about causation, and, as Alicia Juarrero (1999) argues, there is reason to think that our difficulties with differentiating the actions of an agent from mere movements can be traced to an implicit commitment to this "inadequate, 350-year-old model of cause and explanation" (p. 3).

Even before we get whether a different conception of causation can do a better job for describing agency, we can note that this view of matter as inherently inert and changing only in response to the external determination is also incompatible with contemporary physics, which conceptualizes matter not as static but as intrinsically dynamic. Even at its lowest energy state, a particle like an electron will constantly "jiggle around" whatever region of space it is confined to, and the smaller the constraint the faster the jiggling. As Luisi and Capra (2014) describe,

> This tendency of particles to react to confinement with motion implies a fundamental "restlessness" of matter that is characteristic of the atomic world . . . The fact that particles are not isolated entities but wave-like probability patterns implies that they behave in a very peculiar way. To the extent that things can be pictured to be made of smaller constituents—molecules, atoms, and particles—these constituents are in a state of continual motion. Macroscopically, the material objects around us may seem passive and inert, but when we magnify such a "dead" piece of stone or metal, we see that it is full of activity. The closer we look at it, the more restless it appears . . . Modern physics thus pictures matter not at all as passive and inert but as being in a continuous dancing and vibrating motion whose rhythmic patterns are determined by the molecular, atomic, and nuclear configurations. There is stability, but this stability is one of dynamic balance, and the further we advance into matter, the more we need to understand its dynamic nature to understand its patterns. (p. 75)

In modern physics then, the Newtonian order is reversed. There is no answer to what event "caused" an electron to wiggle other than that this is just what electrons do. So, how do we think about causation in a world where change is the default and it is stability that is in need of explanation?

One option is to take causal relations to hold only at the macroscopic level where things at least *appear* to be stable unless perturbed by something else. Such a view often ends up treating causation as useful for explanatory purposes but absent in fundamental physics, or, as Price (1992) puts it, as "anthropocentric, being linked to our perspective as agents." Insofar as "our perspective as agents" is exactly the thing I take to be in need of explanation, this view will not help us here.

Moreover, as Hoffman (2012) argues, it is precisely the scale of the microscopic molecular storm at which life begins its operations, and in terms of which it must be explained. The problem of a living system, as he puts it, is that "without the shaking and rattling of the atoms, life's molecules would be frozen in place, unable to move. Yet, if there were only chaos, there would be no direction, no purpose, to all of this shaking" (p. 21). In chapter 8, I briefly touched on how this molecular chaos may introduce genuine randomness into genetic replication, but of more relevance here is how it interfaces with an organism's metabolic operation. As Godfrey-Smith (2016) describes,

> Metabolic processes in cells occur at a specific spatial scale, the scale measured in nanometers—millionths of a millimeter. They also take place in a particular context, immersed in water. In that context and at that scale, matter behaves differently from how it behaves elsewhere . . . There is unending spontaneous motion that does not need to be powered by anything external. Larger molecules rearrange themselves spontaneously and vibrate, and everything is bombarded by water molecules, with any larger molecule being hit by a water molecule trillions of times per second . . . The way things get done is by biasing tendencies in the storm. (p. 4)

It is because of this dependence on spontaneous molecular motion that Godfrey-Smith suggests that metabolisms could not have arisen at any other scale. An interesting question to look at once we've finished formalizing metabolism in terms of constraint closure is whether such processes could nonetheless extend beyond this scale once they've gotten going. Such a question will be important when it comes to explaining how any teleology we locate in metabolism is to be carried over to minds.

So, in order to make sense of the particular causal regime at work in living systems, both Juarrero (1999) and the theoretical biologist Robert Rosen

(1991, 1999) argue that we need to step out of this Newtonian framework and look back to how causation was expressed in Aristotle's account of the "four causes"—efficient, material, formal and final, of which Newtonianism retains only the first.

Neither Juarrero nor Rosen intends to suggest that we revive the Aristotelean framework wholesale—indeed Juarrero blames Aristotle's "prohibition against self-cause" as much as the "billiard ball" reduction of causation for the problems with our contemporary theories of agency. Rather, their proposal is that, among these various concepts of causation, we can find ones more apt to describe how it works at the molecular level of biological operations. For Rosen, this is formalizable in terms of recursive functions and their variables, in which the material "cause" is a variable whose state is transformed, while the function is the efficient cause that brings this about (though as Moreno & Mossio [2015] note, "formal cause" seems potentially more apt for the latter).

For Juarrero, it is a matter of the more concrete notion of constraints. Constraints can limit and stabilize movement, as when bonds between atoms limit their degrees of freedom and calm them into the apparently inert solid objects around us. Juarrero, however, is more interested in how constraints can not only limit possibilities but enable them.

This idea of a constraint as something that expands the range of possibilities can, at first, seem orthogonal to the meaning of the term. Nonetheless, constraints that "make things happen" precisely by preventing other things from happening, are a pervasive feature of the world and we need not look to the biological to find them. It is by restricting a restless teeming mass of atoms into a small space that the cylinder of a steam engine can perform the work of making a wheel go round, and insofar as the cylinder also remains unchanged during this thermodynamic flow, so it qualifies as a constraint.

Constrained or not, the energy from burning coal would still be transferred. Without constraints, the only way for this to occur is via heat transfer as the hot gas expands and the kinetic energy of these gas molecules is distributed to the surrounding air molecules as they collide. What defines heat, as opposed to work, is that it involves not just a transfer but also a loss of "useful" energy available to do further work. Through heat transfer, energy becomes dispersed into an object's surroundings—corresponding to an increase in entropy. By contrast, work involves "the constrained release of energy into a few degrees of freedom" (Atkins, 1984) such that the "concentration" of energy remains constant—as when energy from thermal expansion is channeled into the raising of a piston, an increase in potential energy that can then be transferred into the movement of a wheel. The nice thing about work is there is always more of it that can be done.

So, the more constrained the energy, the lower the increase in entropy (energy dispersal) and so the greater the amount that is available to perform mechanical work. When indexed to temperature, it is this quantity of constrained energy that makes up thermodynamic free energy, which Friston has sometimes played fast and loose with analogizing to the statistical construct involved in free energy minimization. It is the last thing an organism wants to minimize if it is to continue doing the work of interacting with its environment and generally staying alive. Because there is no such thing as an ideal engine, however, free energy is also lost through heat whenever work is done. Hence even if internally directed metabolic processes were the only work that organisms performed, they would still need continual resupply of energy if they are to continue doing so.

The funny thing about constraints is that the new possibilities they create are not found at the microscopic level of one-to-one interactions between individual chemicals. A catalyst is the paradigmatic example of a constraint, but it does not make reactions happen that would otherwise be impossible; it merely provides a lower-energy route for said reaction to occur, allowing it to happen much more frequently. Sugar oxidation, for instance, is an exergonic reaction, meaning that it happens spontaneously and releases energy, yet confectioners are rarely consumed by unexpected fireballs. This is because sugar oxidation normally happens extremely slowly. Even if you mix it in a glass with an oxidizing agent, like potassium chlorate (used in fireworks and matches), nothing will happen. But the moment you add a drop of sulfuric acid, the two chemicals will instantly react in a violent purple explosion, destroying the glass in which they're contained (Shakhashiri, 1983).

The sulfuric acid doesn't exactly *do* anything, in the traditional sense of event-causation. Rather it is a catalyst that acts as a constraint, remaining unchanged throughout the reaction but accelerating the rate at which it occurs. At the level of the individual molecular reactions, this is just the same thing that would have happened anyway—just much, much faster. At the macro level of glasses, and the people who own them, it's the difference between a beloved crystal tumbler and a pile of glass shards in need of explanation.

In a similar manner, all things being equal, water molecules at the top of a hill will end up at the bottom, dispersing their potential energy in the process. It is only if this flow of energy is constrained by a channel that it will maintain sufficient concentration so as to turn a water wheel on the way down. By making a quantitative change to the spatiotemporal scale of microscopic processes, a constraint can make a *qualitative* difference in the production of a new macroscopic effect.

So, the transfer of energy stored in an object to its surroundings is the norm and will happen spontaneously as the system moves into a more thermodyamically stable state—with a corresponding loss of "useful" free energy via heat transfer and an overall increase in entropy. But for work to be done, for a macroscopic mechanical event to occur with the potential to cause further events in turn, this energy transfer must be constrained. As such, as Kauffman (2000) argues, when explaining why a wheel has been rotated or a glass destroyed, pointing to the transfer of energy seems to focus on the wrong locus of explanation. Energy transfer would happen either way, but the reason work has been done is because this energy transfer was *constrained*.

So, this is the sense in which a constraint can be a cause. The next question we need to answer, in order to explain how we can have a closure of constraints, is how something that is defined partly by its invariance through change can nonetheless be the effect of an ongoing process, on which it depends for its continual regeneration.

10.2.2 Constraint Production

As I have described, there is nothing particularly special or agential about constraints as causes. They are a feature of any machine that channels energy to perform work. In a machine, however, constraints are rarely treated as effects in turn. Once a machine has been created, the constraints that make up its structure can be taken for granted as invariant features that stand outside of the thermodynamic flow. It is in this sense that constraints that are part of a typical physical model are described as "external." And it is for this reason (in addition to their further decomposability into structural parts) that machines are so well modeled in mechanistic terms of a division between fixed equations and parameters versus dependent variables. When we make such a distinction in a dynamical systems model of a machine, this model is picking up a real feature of the target phenomenon, in a way that I will argue it does not in the case of living systems.[2]

Given that Moreno and Mossio (2015) define constraints as things that are unaffected by processes, and which are capable of "harnessing a thermodynamic flow without being subject to that flow" (p. 15), so it may seem as though any constraint-based model of organisms must be unavoidably machine-like. Yet constraint closure is specifically supposed to capture how we can have constraints that do not stand *entirely* outside of the flow of activity in the system (as in a machine) but depend on this activity for their repair or regeneration.

The crucial caveat in making sense of this tension is that the invariance of a constraint holds only "under certain conditions or from a certain point of view" (p. 11). What is meant by this is not that whether or not something is a constraint depends on whether you choose to treat it that way, but rather that a constraint is always defined relative to the processes that it constrains, as shown in figure 10.2. Taking an enzyme (C1): the reaction it catalyzes (A1>B1) has a timescale (τ1) over which it occurs and with respect to that timescale, the enzyme remains invariant. Relative to that process an enzyme is a constraint. Over a longer timescale (τ2), however, this enzyme will degrade and be in need of repair by the process of translating an mRNA sequence into the chain of amino acids that makes up that enzyme (A2>C1), an assembly process that can only happen because it is constrained in turn by ribosomes (C2), which also need to be replenished in turn. And so on and so forth.

So, something that is a fixed constraint over one timescale is nonetheless a stage in a continuing process over another timescale. As the biologist Ludwig von Bertalanffy expressed the same idea, "the old contrast between "structure" and "function" is to be reduced to the relative speed of processes within the organism. Structures are extended, slow processes; functions are transitory, rapid processes" (von Bertalanffy, 1941, p. 251—quoted

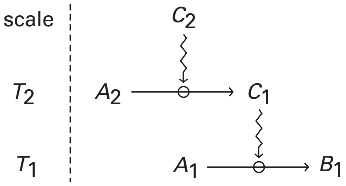


Figure 10.2
An illustration of the dependence of constraints on other constraints from Moreno and Mossio (2015), attributed to Maël Montévil.

in Dupré and Nicholson, 2018). What distinguishes biological constraints from those found in a typical machine is that they are not intrinsically stable things, but extrinsically stabilized processes.

This point can seem a bit odd, given that we have been emphasizing the need for two different levels of causation in characterizing an organism—the need for a distinct role for processes versus constraints in our description of the system. If constraints are really just processes, then what makes constraint closure importantly different from the process closure of Di Paolo and Thompson (2014) and other contemporary enactivists? The answer is that because the latter speaks only of processes and enablement between them it does not explicitly address how more stable processes can serve as invariant constraints with respect to a faster timescale process that they enable specifically *by constraining it.*

As such, a process closure view can lead one to think of the relatively invariant structures that stand above and constrain a particular thermodynamic flow as non-processual entities—as things that are completely external to this flow. In other words, focusing on the closure of processes alone may not mandate the absolute process/structure distinction of a machine, but it is compatible with a perspective in which this distinction is preserved.

The insight of constraint closure is not that constraints are separate from processes, but that in biological systems these constraints are themselves part of the processual network, not fixed structures standing outside of it. By defining the invariance of constraint relative to the process it constrains, Moreno, Mossio, and Montévil's account explains how something can be both invariant over one timescale and changing over another, without making this choice of timescale a subjective decision of some external observer. With regards to mechanistic and dynamical systems models, their account of precarious constraints nicely explains how these can work over one timescale, while failing over a different one. This idea of precarious constraints thus explains how, in biological systems, the distinction between a process and a constraint can be simultaneously real and relational.

Just as there is nothing particularly biological about constraint causation, so there is nothing particularly unique about constraint production either. When an automated digger constrains the flow of energy from a battery so as to channel this into the mechanical work of digging a channel, it is creating a constraint that can, in turn, channel a flow of water down a hill so as to power a water wheel. We might even hook the digger up so that it is powered by this same waterwheel and imagine that the soil quality is poor so that the channel is constantly collapsing and must be continuously re-dug. In this situation, we have a precarious constraint that is necessary to enable the work that maintains that same constraint.

What we don't have here, and what we do have in the organism, is constraint closure. The dependence between the channel being dug and the water flowing through it to power that digging is not direct, but mediated by a variety of other constraints, such as the structures of the waterwheel and the digger, that it plays no part in maintaining and which are stable in their own right. In contrast, to realize closure *every* constraint within the system must not only enable other parts of that system but depend on other parts in turn. Every enzyme, ribosome, membrane, and mRNA strand in a cell is not only a constraint on its metabolic network but also a product of processes that can only occur because of the constraints of that very same network.

To say that an organism is a closed network of constraints does not mean that it doesn't depend on the external environment. Crucially, as with most attempts to formalize the kind of closure distinct to an organism, Moreno and Mossio (2015) emphasize that this runs alongside necessary openness—in this case, to thermodynamic flows of energy and the reactants that fuel these processes (as shown in figure 10.3) In this respect, just like Thompson

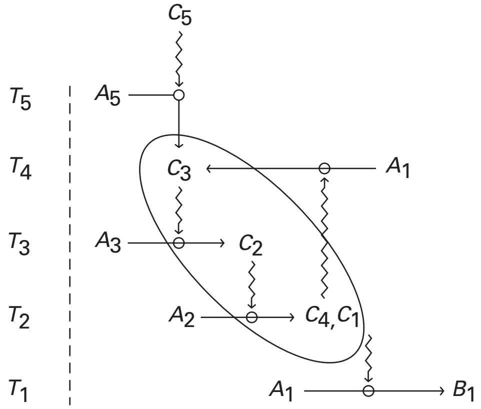


Figure 10.3
An illustration of constraint closure from Moreno and Mossio (2015), attributed to Maël Montévil. C*x* are constraints, A*x* are reactants, and B*x* are products. The wavy lines depict a constraining relationship, and the straight lines depict a process of transformation.

and Di Paolo's (2014) definition of an operationally closed network of processes, a constraint-closed system is both distinguished from, and constitutively dependent on its environment as a resource. What differentiates the constraint closure formulation, however, is an emphasis that it is not only the *activity* the system engages in that is dependent on environmental resources but also the relatively invariant structural constraints that enable this activity, which are themselves similarly precarious and dependent on the environment for the inflows of matter needed to reconstitute them.

Further, constraint closure allows that the system can depend on external constraints to channel these thermodynamic flows *to* it. But any constraint that channels a flow *within* the system—any constraint that operates between one constraint that is a part of a system and another—must also be regenerated by that same system if it is to realize closure. Thus, all machines, even those whose activity depends on its preservation of a precarious constraint, like the Digger-Waterwheel-Channel-system, will fail to meet the criterion of operational closure, insofar as this activity is mediated via a number of non-precarious structures that will remain stable in their own right, independent of the rest of that network and its operations.

So, this is how constraints can cause, depend on other constraints, and come together to realize closure. There is much more to Moreno and Mossio's account in terms of how additional processes of regulation, adaptivity, and evolution lead to the increasing complexity of constraint-closed systems, both within and across generations. This brief account is also highly abstract in relation to the fine-grained empirical details of its realization in the intimidatingly complex network of real metabolisms, which even in their very simplest microbial form enlist several hundred reactions and metabolites. These connections have been elaborated more thoroughly by Stuart Kauffman (1986, 1993) in his work on autocatalytic sets, as a more empirically detailed precursor of Montévil and Mossio (2015) and Moreno and Mossio's (2015) account. There are also other closely related accounts of what Letelier et al. (2011) review more generally as "metabolic closure"—not least Robert Rosen's (1991) M-R systems, which serve as a direct, though in this case more abstract, inspiration for constraint closure.

But it would take another book to explore all of this. For now, the question is what makes constraint closure better than either "closure among processes" or steady-state ESIA-closure for defining life? And does it do any better at locating the basic ingredients for a real notion of intentionality and purposiveness from amid natural processes?

10.3 Constraint Closure as a Theory of Living Systems

The most basic challenge of an account of living systems is to obtain the right balance of specificity and generality. While there may be some disagreement on the status of border cases such as the cryptobiotic organism, the virus, the seed, or some, as yet unrealized, sophisticated AI, the criteria proposed should at least succeed in including those things uncontroversially regarded as living and in excluding those cases which no one not already in the grip of some particularly contorted theoretical commitments would mistake for being alive.

Moreover, a successful account also needs to be able to deal with the possibility for life to have evolved in different ways, under different conditions. Simply tying life to some particular chemical manifestation, even if this is shared by all known instances on Earth, would be the most obvious failure to describe “not just life as we know it, but life as it can be” (Langton, 1989). A central advantage of relational, or “operational,” accounts like autopoiesis is that they say nothing about the particular chemical compounds involved, only the organization they need to be able to realize in terms of the relations between them (Fleischaker, 1990).

Still, in focusing on molecular metabolism, autopoiesis is insufficiently general to describe life at the multicellular and sensorimotor levels. For this, Maturana and Varela initially proposed the idea of second-order autopoietic systems, composed of autopoietic subunits (1987, pp. 88–89). This, as Thompson (2007, pp. 105–106) describes, is not quite satisfactory. How exactly do the autopoietic processes cohere in an organismal whole? And is it the autopoietic subunits or the overall organization of the multicellular organism in virtue of which it is to count among the living?

Because the bioenactive approach seeks to characterize the minds of multicellular creatures, not just life in its most basic single-celled form, so the more general notion of autonomy has taken on a more important role than autopoiesis. Yet in moving away from the single-celled level to focus on the closure of precarious processes, such characterizations have relaxed more than just the restriction of autopoiesis to molecular synthesis. Specifically, as argued in the earlier part of this chapter, this account loses sight of the relation between these processes and their products, between the activity of the system and the body it regenerates, which was core to autopoiesis. As a result, the present characterization of autonomy as process-closure is now too general to distinguish the living, given how it may be instantiated in nonliving systems like the hydrological cycle or the robot-swingball set.

Process-closure still does better than the FEP's steady-state ESIA cycles, which somehow managed not only to be too general—in applying to nonliving systems—but also too specific—in specifying necessary requirements that living systems themselves do not meet. Both ESIA-closure and process-closure fail for a shared, and more basic reason however, that they go too far in their abstraction—not only disassociating life from any particular chemical realization but also from its unique thermodynamic status. In autopoiesis, these energetic and material requirements were arguably at least implicit and only in need of further elaboration (Fleischaker, 1988). In contrast, by placing the structure of constraints entirely outside of the thermodynamic flow of the system to focus only on processes, so contemporary enactivist accounts of autonomy succeed in providing a substrate neutral, multiscale characterization of closure—but only at the expense of erasing this crucial distinction between precarious, non-equilibrium constraints versus intrinsically stable ones (Ruiz-Mirazo & Moreno, 2004; Bickhard, 2000).

As such, when seeking to account for the distinction between living systems and other circular organizations, these accounts can only resort to either the requirement of autopoietic sub-components or, as Ashby did, to arguing that such a distinction just comes down to the degree of complexity of the mechanisms involved. Thus, Di Paolo (2005) proposes adaptivity as an additional requirement for life and, Di Paolo, Buhrmann and Barandiaran (2017) propose this, along with further supplementations as being necessary for agency.

I think Di Paolo, Buhrmann, and Barandiaran (2017) are correct that such properties are required for agency, and perhaps for life too. However, I take these to be intrinsically linked to a more robust notion of self-production they provide. Without a criterion capable of distinguishing autonomous agents from automated machines, the distinction between life and non-life, the coupled-pendulums and the bacteria, becomes a matter of the degree of adaptivity or complexity of some network of processes. Life itself becomes a purely abstract phenomenon that might as easily be instantiated in a computer as a chemical network.

In this respect, such organizational, or relational, definitions can be contrasted with work that has focused specifically on the thermodynamics of life, and on living systems as far from equilibrium dissipative structures, such as a candle flame, a convection roll, or a tornado, that are constitutively dependent on ongoing flows of matter and energy for their continuation. (Schrödinger, 1951; Bertalanffy, 1968; Nicolis & Prigogine, 1977; Fleischaker, 1988; Juarrero, 1999; Christensen & Hooker, 2000; Christensen & Bickhard, 2002; Collier, 2004, 2008). Such dissipative structures seem to manifest the needful freedom in their relationship to matter that Hans Jonas (1953, 2001/1966) took as the hallmark of the immanent teleology unique to living systems. Yet while the dissipative structure and living organisms may share this precarious form of existence, it seems just as problematic to ascribe intentionality and vitality to tornados, convection rolls, or a candle flame as it would be to ascribe such properties to a computer system or to coupled pendulums.

Accounts of life seem caught between the metaphor of the candle flame and that of the computer (Keller, 2008, 2009). Neither the organizational nor thermodynamic properties of a living system alone suffice to capture what makes an organism them distinct from both.[3]

As Fleischaker (1990) describes,

> It is not new, of course, to point out that life requires energy to drive its processes of production. Nor can it be claimed that life is alone in requiring energy for the integrity of its internal organization: fluid or gaseous convective systems utilize heat-driven density gradients to that same end. These complex dynamical systems are non-living, but they, too, transform energy from the environment in maintaining themselves at a distance from equilibrium, and they hold energy in non-linear relationships among system components, that is, in circular (self-amplifying) relationships in which effects become cause (Swenson, 1989). What is unique to living systems is the organized coupling of energetic and material interactions in a single network of processes whose outcome is the production of all system components, including the constituents of its membraneous boundary structure. (p. 128)

The strength of constraint closure then is how it combines the insights of both organizational and thermodynamic approaches to living systems. As a relational notion, a precarious constraint is more general than the notion of a polypeptide chain, or an enzyme—yet something's ability to stand in the relevant relations is highly constrained by its material properties. Namely, this material must not only be capable of channeling a thermodynamic flow to enable a particular process, but it must be both invariant throughout the timescale on which this process occurs, and unstable on the larger timescale of a slower flow to which it is also subject.

It is because of these particular thermodynamic requirements that the notion of a precarious constraint, while medium-variable, is not medium-independent, and why metabolism is not a purely formal property that can be literally realized in a computer model, as Boden (1999) argued. We may describe and interpret parts of a computer simulation as representing "flows of energy" and "precarious constraints upon those flows," but so long as the bits of silicon and metal that realize this simulation are not actually dependent on the energy that flows through them, so there will always be a gap between even the most detailed simulation and a real metabolic network.

In contrast, a dissipative system has the right thermodynamic properties to literally be a precarious constraint. Yet in such cases of spontaneous "self-organization," there is only mutual dependence between a single constraint and its process of regeneration, and so they do not realize the organizational properties of constraint closure proper to the biological, in which "constraints are not able to achieve self-maintenance individually or locally: each of them exists insofar as it contributes to maintaining the whole organization of constraints that, in turn, maintains (at least some of) its own boundary conditions" (Moreno & Mossio, 2015, p. 17).

This is important for explaining why dissipative systems both spontaneously emerge and then vanish almost as quickly in a way that living systems do not. Where self-organizing systems can arise whenever the boundary conditions are appropriate—for instance when heating from below creates the appropriate temperature differential for the movement of individual molecules between the top and bottom of a liquid to spontaneously organize into the coherent rotating cells characteristic of Rayleigh-Bénard convection—these boundary conditions remain outside the influence of the dissipative system itself. As such, the system has no influence in sustaining them and will disintegrate as soon as those external constraints falter—either due to depletion of the energy source that sustained them or due to external perturbation. A simple self-organizing system may arise spontaneously, but once it does, it cannot influence its environment to support its continued existence (Ruiz-Mirazo & Moreno, 2004).

So, as Moreno and Mossio argue, the distinction between a simple "self-organizing" process-constraint loop and constraint closure is not an arbitrary matter of the number of parts of the system that are acting as constraints. What distinguishes constraint-closed systems from self-organizing ones is not just that there is a greater number of constraints, but that there is a hierarchy of constraints operating at distinct levels and over different timescales—some of which serve as boundary conditions for the possibility of closure between the others.

In this regard, we can point to the significance of a cellular membrane, as an example of this separation. For Ruiz-Mirazo and Moreno, the significance of the membrane for the cell is not spatial as Maturana and Varela (1973/1980) describe it in marking out "the topological domain of its realization as a network." Rather, they argue, the key feature of a membrane is that it is not just another constraint within the metabolic network of enzymes, but that it is a higher-level, slower timescale constraint that preserves the conditions for that network's operation, and is regenerated by it in turn. Even in more complex systems such as autocatalytic cycles, which may self organize under quite specific conditions, these cycles still involve only the timescale of the catalysts and of the processes they catalyze (Virgo & Ikegami, 2013; Virgo et al., 2014). The network has no influence on the higher-level constraints that make this network possible—neither on channeling the flow of reactants into itself, nor on the container in which said network is housed.

It is because biological systems require the coordination between at least two different timescales of constraints that they do not just emerge spontaneously in the manner of a "one-level" self-organizing system.

In biological systems, as Moreno and Ruiz-Mirazo (1999) argue,

> one has to take into account not only the amount of time that a reaction—or some other process—requires in order to be carried out, but also (and most especially) the time it needs in relation to other reactions with which it could become coupled. In other words, metabolism necessarily requires the synchronisation of a whole set of biophysicochemical processes. (1999, p. 51)

What makes this synchronization across different timescales so important is that the activity of said system depends on spontaneous reactions, where parts of the system release energy through their degradation (exergonic reactions), which is then channeled into the energy-absorbing (endergonic) work of self-construction and repair—the latter processes being unable to occur without the energy released by the former. The problem is that, left alone, these spontaneous reactions may happen too slowly for the constraints upon the energy they release to have any macroscopic effect. Imagine attempting to constrain sugar within a piston and expecting its slow oxidation to drive an engine. As described earlier in this chapter, we can accelerate this energy release via a catalyst (which is itself just another constraint), and it is through the use of such enzymes that organismal metabolisms are able to extract energy from glucose in order to power the work of their self-repair. But these catalysts themselves only exist *because* of that regenerative and reparative work. In a living system, there will be redundancy within an individual constrained process, for example, an excess of enzymes or energy stores, but if this constrained process fails altogether, then it will bring down the rest of the exergonic-endergonic couplings making up that organization and the whole thing will start to fall apart.

So, constraint closure requires a number of independent puzzle pieces to come up together in just the right way at just the right time. Such a complex synchronization of energy-releasing and energy-absorbing processes, coordinated with respect to the various different timescales at which these reactions occur, does not just spontaneously emerge "all at once" as in a self-organizing system—hence why tornados, but not tardigrades, can arise when the weather conditions are fortuitous.

10.3.1 Constraint Closure and Adaptivity

Insofar as there is a separation between timescales, so there is already the possibility for a minimal form of responsiveness to perturbation built into a constraint-closed system. Perturbations to constraints operating at one timescale can activate another constraint at a different timescale to compensate for these disruptions. One example, suggested by Ruiz-Mirazo and Mavelli (2007), would be how the production rate of the metabolic network inside a cell raises its osmotic pressure, thereby altering the permeability of the membrane, increasing the rate at which waste products are channeled out of the cell, and so bringing that pressure back into balance.

While this form of what Moreno and Mossio call "stability" is not entailed by constraint closure, it does not require any additional mechanism—only the requirement that these constraints are collectively robust to some degree of perturbation via their modulation of each other. It is hard to see how the delicate synchronization between the various processes and constraints of a constraint-closed system could persist over any significant duration without at least some robustness to disruptions.

Constitutive stability, as Mossio and Moreno (2015) describe, is a conservative process that preserves the *same* organization throughout disruption. In this regard, it might be characterized in terms of the homeostatic logic of free energy minimization. For developmental changes that lead to increases in the complexity of a system's organization, however, we need a further level of second-order constraints—which they call "regulatory constraints." These second-order constraints are defined by their being dormant over some timescale, during which they do not serve as necessary constituents of the constraint-closed system. As such, these dormant constraints may change without instantly breaking the constraint-closure that keeps the organism alive.

One example would be repressed genes. Because these repressed genes are typically not active participants in the constraint-closed organization, they may mutate, or be altered, without immediately destroying this closure. When this new variation is subsequently activated, it may result in the production of a constraint that can synchronize with the overall set of couplings making up the organism—as when the insertion and subsequent activation of the lac operon genes in some prehistoric *E. coli* led it to replace glucose-metabolizing constraints with lactose-metabolizing ones. In such a case, the organization of a living system may mutate into a new form without any break in its continuity of constraint production.

The nervous system could also be construed in terms of an even more decoupled set of constraints. Repressed genes like the lac operon will, when activated, enter directly into the new constraint-closed organization by producing constraints that channel energy into productive work. But even when energy courses through a neuron, it is not directly put to metabolic ends. Instead, the energy entering at the sensory periphery is channeled into coordinating the multicellular organism's motor system, to seek out the sources of energy that the metabolic network needs and to avoid those things that would threaten it. This double decoupling, as Moreno and Mossio suggest, is what affords a much greater degree of adaptivity and plasticity to neurally equipped creatures.

Importantly, however, while these decoupled subsystems are not mandated by the notion of constraint closure in itself, they remain part of the constraint-closed organization, even when decoupled from its ongoing constitution. Dormant genes may not be contributing to an organism's metabolism over the time period of their dormancy, but their existence depends on their having the potential to make a beneficial contribution under some conditions. If a particular neural assembly does not successfully channel energy into behavior that supports the metabolic system, it is liable to be rewired into a more beneficial format. So, such regulatory constraints are *enabling,* insofar as they facilitate the transition between different constraint-closed organizations, rather than the movement of energy through different production processes within an organization. They are also dependent, insofar as their preservation depends on their success in doing so. As such, Mossio and Moreno argue that "regulatory constraints are subject to a *second-order closure* between both themselves and the whole *set* of organisations among which they govern the transitions" (p. 35).

So constitutive stability and regulation are not intrinsic to constraint closure, any more than adaptivity was to autopoiesis. They are, however, implied by the requirements of preserving constraint closure in a world where things change. Once we have regulation, we not only have an explanation of how something as delicately balanced as a constraint-closed system could persist in such a world but also (the start of) an explanation for how these changes could lead it to evolve into the vast variety and complexity of metabolic organizations that we see today.

As Moreno and Mossio (2015) put it,

> Biological organisation must be able to handle variations, and then conserve closure, otherwise it would be extremely fragile and its realisations in the natural world would hardly move beyond a very low level of organisational complexity. Any perturbation would be more likely to drive the system to disruption than to result in an increase of complexity. What is then required for biological organisation not only to remain stable in the face of perturbations, but also be able to increase its complexity? The answer is, we submit, regulation. Biological autonomy requires regulated closure. (p. 30)

Thus, we can agree with Di Paolo (2005) and Di Paolo et al. (2017) that adaptivity is necessary for a biological system. Yet, by taking this to be grounded in something more robust than operational closure among precarious processes, we secure a more plausible distinction in kind between autonomous and non-autonomous systems. This problem with operational closure's application to nonliving systems is that it invites the idea that whether something is a living or autonomous system will depend on something like the degree of complexity and adaptivity of this network of mutually dependent processes. Unlike process closure, constraint closure allows us to explain how this capacity of adaptivity, which may be graded, is tied to a more fundamental difference in kind.

So, constraint closure provides a way of describing the relational logic of a living system that is "geared in" to the energy flows that must be harnessed for its realization. In doing so, it identifies a qualitative difference between the existence of an organism versus that of either an intrinsically stable machine-substance or a spontaneously emerging dissipative system. Moreover, constraint closure gives us an account of the kind of relations of production that must be instantiated at every point in the organism's development, without requiring the organism be fixed to any one particular invariant organization to realize those relations. As such, it allows for the possibility of adaptable regulatory mechanisms that can alter the organization of our constraint-closed system without destroying this ongoing process of self-production.

I have argued that constraint closure provides a means to describe the dependence of organic structures on the processes that support them. Its description of this dependence improves on both the FEP's steady-state ESIA cycles and the bioenactive notion of operational closure in its ability to capture the difference between living systems and nonliving systems. Still, the fact that there is such a difference in kind does not entail that this distinction renders living systems, and living systems alone, *agents*. So, how does the constraint-closed organization of living systems underpin their capability to engage in actions that are intentionally directed and normatively evaluable?

11 Constraint Closure as the Basis of Intentionality

This book began with the claim that the foundational question for the enactive approach is the question of what it is for something to be intentional in the sense of an "act having directedness" toward some norm that it might fail to satisfy. It is this question that the notion of autonomy is supposed to provide a solution to. The issue I took with many theories typically lumped together with the enactive approach, such as sensorimotor enactivism, radical enactivism, or embodied cognitive science more generally, is that they do not take autonomy as a central question. Rather than seeking to account for the normativity of actions they either take it for granted, reject it altogether, or treat it as something that cannot be derived from the properties of individual agents. Without normativity, what we have is not so much enactivism as mere (sensor-guided) movementism.

To be more than mere movementism, enactivism not only needs an account of how normativity gets into the world in general but, specifically, an explanation of how this normativity gets into the actual movements of an organism. An act, as Davidson (1963) emphasizes, is not just a movement that it's consistent with, or that satisfies a normative requirement, but one that is caused in the right way by this normative requirement. Thus, a theory of the difference between directed actions and mere movements must not only explain what a purpose (or reason, goal, or norm is), as opposed to a mere cause, but also what it is for some movement to be performed *for* one, as opposed to merely being interpretable in terms of a reason (Wittgenstein, 1953/2010)—as when we say silly things like, "The book stayed perfectly still for the purpose of remaining unnoticed," or, "The pendulum returned to rest for the reason of minimizing its free energy."

It is only by answering both these questions, in order to distinguish between acts and mere motions, that we can obtain a robust sense of intentional-directedness that differentiates the actions of an agent (or at least a proto-agent) from the motions of a machine. The "reason for an action," as Hurley (1998) argued, is no more of an unproblematic given than the "content of a perception." Rejecting reconstruction in favor of the practical normativity of actions is no solution if we still lack a foundation for these normative attributions. Without this, the enactivist is just as vulnerable as the reconstructivist to the objection that their "intentional acts" are nothing more than instrumentally useful abstractions, whose validity is relative to our own explanatory perspective.

As I'll describe in this chapter, I believe that an adequate account of biological autonomy can explain how normative requirements get their claws into actual actions. But before we dive in, I should note that in the analytic tradition, where this question has been extensively discussed, the standard interpretation of a reason has been in terms of propositional attitudes, such as beliefs and desires. An action is taken to be caused by a reason if it results from a non-deviant causal chain corresponding to a valid inference from the content of these propositions to the action that is entailed by them. Acting for reasons becomes tied to the possession of fully fledged conceptual capacities, language, and the capacity to use these to justify your actions to others.

This is not the approach I wish to take due to a number of considerations. Firstly, as a sympathizer with the bioenactive approach, I take the possibility of having intentional attitudes to be the thing we should be in the business of explaining, not something presumed at the outset. Secondly, as someone influenced by the phenomenological accounts of intentionality, discussed in section 1.1.2, I am skeptical that propositional attitude psychology will prove to be the best way to talk about our intentional attitudes and how they guide action. As a result, I am convinced by Hurley's (2003) arguments that to understand practical rationality in the model of "theoretical rationality with practical content" is to overintellectualize the mind by mistakenly prioritizing epistemic "reasons for belief" over practical reasons for action.

Independently of all the above commitments, I am also unconvinced that this explicitly inferential account is sufficient to overcome Wittgenstein's rule-following dilemma. As Boghossion (2014) points out, inference itself is a normatively guided action and the attempt to explain normatively guided actions in terms of inferential processes collapses into a regress. Rather than abandoning talk of "reasons" to the traditional rationalists, then, I want to keep the term while showing that understanding it differently can provide a solution to the rationalist's troubles. Reason is, after all, a particularly useful term for its ambiguity between causal and normative factors and many of the same problems with making sense of it can be developed within the pre-linguistic sphere of organismic activities.

So, in speaking of the normativity of organisms in terms of reasons, I do not mean to overintellectualize the organism with the baggage of propositional attitudes and explicit inferences, as Fulda (2017) criticizes. Rather, in the spirit of Hurley's (2003) proposal, the intention is to deintellectualize rationality. Moreover, my suspicion is that the rationality observed in social and linguistic practices will emerge as a special case, or a complexification, of the kind of normative practices that first arise in more basic organic forms. What unites all these activities as reason-guided, what distinguishes them from rocks rolling toward a valley floor or rivers running to the sea, is that they occur because of some standard that they *should* satisfy but may nonetheless genuinely fail to achieve. It is this exceptional property that the following section aims to sketch an account of.

11.1 Organisms as Networks of Reasons

The idea that we can distinguish between something being done *for* a reason, as opposed to merely being describable in terms of a reason, implies that there is a fact of the matter about whether some particular reason is responsible for why an action happened. Yet reasons are also often opposed to causes in that where the former bear a normative relation to what they are supposed to bring about, which can fail to be satisfied, in the latter case the relation between the cause and the effect is one of necessity governed by exceptionless natural laws. As Kenny (1989) describes this,

> One important difference between the explanatory power of reasons and the operation of causes is this. If there is present a perfectly adequate cause for an effect, then the effect cannot but follow: for a cause—at least on the determinist's view of the matter—is a sufficient antecedent condition for the effect, and if an effect does not follow when an alleged cause is present we know the cause is not a genuine one. On the other hand, there may be a perfectly adequate reason for performing an action and yet the action may not ensue, without this fact casting any doubt on the adequacy of the reason. (p. 145)

So, when I say, "Hmm, that mixture of sugar, potassium chlorate, and sulfuric acid shouldn't have broken the glass," I am merely revealing my ignorance of the force this reaction would exert on its container. When my partner responds, “*You* shouldn’t have broken my glass,” he’s saying that there was something wrong or incorrect about the action that led to it breaking, and he will maintain that conviction irrespective of whatever additional scientific details are offered in response. What distinguishes the normativity of reason from the necessity of cause is that, in the former case alone, there is supposed to be a genuine possibility for what *ought* to happen and what actually does happen to come apart. Thus, as Longo and Montévil (2013) describe the role of constraint closure and autonomy in biological explanation:

> A river never goes wrong and we know why: it will follow a geodetics. An onto- or phylogenetic trajectory may go wrong, actually most of the time it goes wrong. We are trying to theoretically understand “how it goes,” between causes and enablement. (p. 16)

It is because normative relations between reason and actions appear incommensurable with a deterministic cause-and-effect universe that reasons or purposes are sometimes either accepted as nonnatural (Parfit, 2006, 2011; Enoch, 2011; Scanlon, 2014), or rejected as nonexistent (Henderson, 2002, 2010). To provide a naturalistic account of how there can be reasons that could genuinely be responsible for actions, we need to explain how something can combine both the force of a law with the possibility of its failure.

It is this tension expressed in what Weber and Varela (2002) describe as the Jonasian antinomies of “freedom and necessity, autonomy and dependence.” The same apparent conflict that is found in Kant’s original formulation of freedom as autonomy (literally, “self-*law*”) in terms of “the will’s property of being a law to itself” such that “a free will and a will under moral laws are one and the same” (Kant, 2008/1785, pp. 446–447). It is arguably a tension better captured in the earlier statement that inspired Kant, Rousseau’s assertion that “the impulsion of mere appetite is slavery, and obedience to the law one has prescribed to oneself is freedom” (Rousseau, 2018/1762, 56). The notion of autonomy thus encapsulates the question of how we reconcile the freedom and necessity that defines normativity—the question of how a law can be at once something contingent or optional, such that it is dependent on the autonomous system that prescribes it to itself, and yet non-arbitrary and binding such that this autonomous system may nonetheless be subjugated to it.

Arguably the most prominent development of this idea within Anglo-American philosophy is the rationalism of Nagel (1986, 2012), Brandom

(1979, 1994), McDowell (1994), Korsgaard (1996), and others who, building on Sellars (1956), point to our distinctive conceptual and linguistic abilities and how the uniquely human discursive practices they enable place us in the "game of giving and asking for reasons" within which participants construct laws that then hold for those participants in turn (Sellars, 1956). For such accounts, autonomy is a matter of this sense in which humans alone are rational beings, capable of communication, deliberation, reflection, and the self-conscious recognition of particular evaluative standards as applying to both ourselves and to others.

As Jebari (2019) argues, "The prevailing attitude is that rationalist approaches to ethics are essentially unworkable from within a scientific context and must be abandoned as part of the naturalizing project in ethics" (p. 1). This attitude is not only held by rationalism's critics. Following the Humean proscription against deriving an ought from an is, many rationalists have also defended the independence of this space of reasons from the realm of science and its laws, which is argued to lack the conceptual tools for a description of normativity.

For this reason, and insofar as such accounts are also committed to making rationality a distinctively human capacity, rationalism appears diametrically opposed to the bioenactivist project of naturalizing normativity via the capacity that humans share with other forms of life. Yet, Jebari (2019) suggests that both these commitments are inessential, and the conflict between the rationalists on the one hand, versus naturalism (and perhaps nonhuman normativity) on the other, is unnecessary. What is important to rationalism, he argues, is not an autonomous domain of human rationality per se, but the attempt to use this to derive a concept of normative facts as being out "there anyway, whether or not [our] eyes are opened to them" (McDowell, 1994, p. 91), without reifying these norms as eternal essences that are "constituted in splendid isolation from anything merely human" (McDowell, 1994, p. 92).

Furthermore, in a similar vein to Juarrero, Jebari argues that the belief that this desire for objectification-without-reification cannot be incorporated within scientific naturalism, or physicalism, stems not from the intrinsic nature of scientific explanation itself but from the philosopher's impoverished concept of what scientific explanation and the physical are. Once we recognize the importance of constraints, he argues, we can see how a scientific worldview has the resources to describe normativity as an objective feature of the natural world. As he describes,

> This construal of the rationalist position also provides a way to satisfy the objectification-without-reification constraint. For, on this approach, whether a normative standard applies to an agent is not generally a function of the agent's attitudes; rather it is a function of (1) the overall structure of the social system and (2) the agent's position in that social system. Normative requirements are thus constituted by structures largely external to the agent, and an agent can do better or worse at recognizing and responding to the requirements that in fact apply to her. Nevertheless, the reality of such requirements does not entail Platonism, since such requirements emerge from perfectly natural social-systemic processes. (p. 15)

So, for Jebari, it is because these constraints are both constructed and realized by social systems that they are, unlike necessary laws of nature, contingent principles that may be violated. But it is because a society is nonetheless a real structure, capable of limiting the behavior of its members, that these constraints are both objective and naturalistic features of the world. Thus, as he explains, "People's behavior will both explain and be explained by these constraints, yielding an overall picture in which people's actions operate in both a norm-guided and norm-constituting capacity, often at the same time" (p. 16).

Still, while Jebari emphasizes the relevant constraints as being those that are constructed, makes reference to the work of Moreno, Mossio, Kaufmann, and others, and discusses how constraints must be organized so as to maintain the system as a whole, his focus remains on their *social* construction. As a result, his account of how these constraints are constructed remains at a relatively high level, in terms of patterns of action and interaction between agents. This sense in which constraints may emerge from the dynamics of a system is much more general than the specific notion of thermodynamically precarious constraints that are sustained only by channeling energy into work. It is this more specific notion of circular constraint construction that I take to be essential in accounting for what makes certain types of constraints distinctively normative.

Constraints are everywhere. Constraints that we have constructed are pretty common too, from the machines we make to the canals we build, but the manner in which steam engine is forced to move when coal is burned or a canal prevents me from walking somewhere, does not seem to me to be a normative affair but a straightforwardly physical one. It is true that these constraints are contingent in the sense that they need not hold. There is no inevitable exceptionless law that matter must form into engines, burn coal, and power steam trains. We didn't have to construct that steam engine, and we could destroy it. There is also an indirect sense in which these constraints are dependent on the activity they enable for their continuation. If the steam engine does a bad job of turning thermal energy into motion, we might melt it down for candlestick holders. Yet, this risk is not inherent to the steam engine itself. We might not destroy it but place it in a museum for schoolchildren to goggle at, where its structure can persist indefinitely without actually constraining anything.

This steam engine *can* constrain energy flows to perform work, but there is no necessity to it doing so. Its constraints are only of the conditional form: *if* there is a flow of energy, then raise a piston. The raising of the piston does not have the force of any sort of necessity, and there is no reason, no need, for the steam engine to do anything at all. In contrast, a constraint-closed organism must be constantly operative because it is only by constraining energy flows that it can enable the regenerative work without which the set of constraints that realize it would degrade—irrespective of what external agents like us might choose to do about the situation. As Nicholson (2018) describes,

> This ongoing self-producing activity is not optional—not undergoing constant metabolic regeneration is not a possibility. The thermodynamically grounded fact that organisms need to keep acting in order to keep existing helps to account for the emergence of a rudimentary form of normativity in nature (cf. Mossio et al., 2009; Christensen, 2012). It is because its existence depends on its own activity that an organism must act in accordance to the operational norms that enable it to persist through time. If the organism stops following these norms, it ceases to exist. What this means is that it is in principle possible to objectively specify what is intrinsically "good" or "bad" for an organism (that is to say, what is and what is not in an organism's "interest") by evaluating its activities according to the contribution they make towards the preservation of its organization in far-from-equilibrium conditions. (p. 154)

Such precarious constraints need to enable regenerative work to continue to exist—as long as they exist, regenerative work *must* be done. In this respect, they have a kind of necessity baked into their existence that ordinary constraints do not. Yet they may also fail. They may not receive the energy they need to do this work. But if they fail, they fall apart. A constraint-closed system needs to enable its own existence, but it is also free to cease to exist. In this sense, a constraint-closed system is itself a "self-given law."

Even before we get to any uniquely human capacities of conceptualization and communication, the constraint closure of organisms already provides the means to break apart the "ought" from the "is." To say this particular *E. coli* bacterium ought to avoid ethanol is not a human projection but an existential imperative derived from the bacterium's self-producing organization. What it is to be that bacterium is to be something that cannot exist in a highly concentrated ethanol solution. Nonetheless, that particular bacterium might still fail to comply with this existential constraint, and thus it will cease to be a bacterium any longer.

This is not to say the demise of a single constraint is the death of the organism, or that the norms they realize cannot change. I argued that what gives the organism continuity over time is not one particular constraint-closed organization but an unbroken relationship of self-production throughout its various organizational stages. For one particular *E. coli* at one particular point, it can be an objective fact that it cannot exist in a highly concentrated ethanol solution, but this *E. coli* might be engineered to survive on ethanol (Cao et al., 2020). From that point on, "avoid ethanol" is no longer a norm that it must follow. Changeability does not make norms arbitrary or subjective, however, and a fact need not be eternally true to be objective. Prior to the development of ethanol resistance, the need to avoid ethanol was an objective fact about the particular organization of a particular *E. coli*. Post insertion of the relevant genes, this bacterium takes on a new organization such that this norm no longer holds.

Because particular individuals can have variability in their organization, so there is no generic fact about whether lactose is "good" for *E. coli* in general. The goodness of lactose depends on whether the specific individual *E. coli* in question has incorporated the lac operon genes into its overall organization, allowing it to channel the energy contained in lactose into its self-production. On the other hand, there is no absolute fact about whether the development of lac operon genes will be good for that particular *E. coli* bacterium either. This depends on facts external to that particular bacterium, namely whether its environment contains a source of lactose that its constraints can channel the energy from.

The only restriction governing these changes in organization is that any new constraints must synchronize with the rest of the exergonic-endogenic couplings making up the organism in the same manner as the one that it replaces, such that the new organization remains a coherent self-producing whole. As far as any particular organism is concerned, the requirement of ongoing production, in whatever organizational guise that might take, is the only ultimate norm that cannot change.

As argued at the end of chapter 9, this means that the description of biological systems must take on a different form from that of ordinary constrained physical systems—in which the invariant constraints and resulting dynamics can be specified at the outset, and do not depend on either historical change or their interactions with the external environment. Thus, the theoretical biologist and the rationalist can agree that the explanation of norm-governed systems is distinct from the kind of explanations given in ordinary physical models. Yet this does not mean that our account of normativity must be nonnatural, only that these particular types of modeling practices fail to encompass all the idiosyncrasies of nature. Naturalistic is a slippery term and I don't want to get into policing its borders here, but insofar as the basic materials of constraint closure involve nothing more mysterious than energy flows and constraints on them—language that would legitimately be found between the pages of *Physical Review*—so it should qualify as naturalistic.

Jebari's account is useful in showing how the generality of the concept of a constraint means that it can be applied to social organizations as well as biological ones—making it apt for scaling the normativity found in molecular metabolism up to the different levels at which normativity might be constituted. But if societal constraints are normative, I propose, it is because societies are like organisms. To understand what makes them so, we should look first to simpler forms of organization in the biological realm, to see how the intrinsic precarity of certain types of constraints place them in the more fundamental game of giving and asking for thermodynamic free energy.

This is the sense in which I take organisms to be "natural purposes" in Kant's terminology. But to be agents, or at least proto-agents, they need not only *be* purposes, they also need to act *for* those purposes. Thus, Korsgaard (2018a) allows that animals can have value and that things can be good or bad for them while rejecting the idea that they are necessarily autonomous systems who are capable of acting for the purpose of obtaining those goods. For Korsgaard, this latter attribute requires a thick Kantian type of rationality as a capacity of self-conscious reflection that makes a human being alone capable of "knowing that an evaluative standard applies to your conduct, that there is a way you should act or ought to act or that it is good or correct to act, and being motivated in part by that awareness" (Korsgaard, 2018b, p. 5).

This view of what it means to be rational, or to act *for* a reason, is not particularly helpful to an enactivist, for whom it is exactly these capacities of being conscious, knowing, or aware, that we want to use an account of autonomy to try to explain. But there is, I will now argue, a much less demanding and more naturalistically grounded sense in which we can argue that living systems are not only reasons but reasons that cause the very actions that they are reasons for.

11.2 Organisms as Causes of Their Own Activity

To say that an organism is both the reason for and the cause of its action amounts to saying that an organism is a cause of itself—an idea that, as Juarrero (1999) argues, is impossible within the Newtonian framework of event-based causation, where every change must be caused by some other event. If this were an accurate description of the natural world, then to say the movement of a system cannot be traced back to through a series of events to some other causes external to it (and, in principle all the way back to the unexplained origin of the universe) would indeed be a supernatural claim (see Juarrero 1999, chapter 1 in particular). As we have seen, however, the natural world is not, at its basic level, Newtonian. Energy will dissipate and entropy will increase with or without external prompting. The macroscopic effects of this dissipation are not determined by some prior external change, but by the presence, or absence, of invariant constraints on those flows of energy.

The idea of constrained energy gives us a different approach to causation, but it does not yet give us self-cause. Machines also operate by constraining the spontaneous release of energy from an unstable reactant, that is, fuel, in order to perform work. But organisms are different from machines in two ways. Firstly, as described, the work that they do involves the reproduction of their own constraints, which would disintegrate without the regenerative activity that they enable. Secondly, organisms not only regenerate themselves; they also consume themselves. In machines, there must be a separation between the reactant that releases energy and the constraints that channel it. There is also energy contained within a machine's structural parts, and this will slowly spread into the environment as those parts rust. But the machine has no means to channel this spontaneous energy-releasing reaction into the work of rebuilding itself. The breakdown of a machine's structure can only ever lead to energy dissipation and destruction.

So, when an engine runs out of its fuel supply, it ceases to operate. But precisely because organisms need to rebuild their component parts, they are also free to accelerate their break-down, via catalysis, and to channel this release of energy through other constraints in order to power further activity directed toward the repair of these, and other, parts of itself. While in Moreno and Mossio's (2015) diagrams of constraint closure, constraints are only depicted as the output of a process, their account also allows that a constraint may also degrade, ceasing to constrain and becoming instead the reactant for another process.

Crucially, this gives us a sense in which organisms are *intrinsically active*, rather than just responsive to perturbation as were Ashby's "sleeping machines" or Friston's free energy minimizers. Like a leaf in the wind, the latter may appear quite animated if their environment is disruptive enough, but their intrinsic dynamics only drive them back toward stasis. In contrast, when an organism is deprived of any input of energy it will continue to operate, first by catalyzing the breakdown of non-constraining energy stores, then via the breakdown parts of the structure by which it operates—such as muscle tissue for amino acids (Steinhauser et al., 2018). Systems that channel energy released from their own dissipation in order to rebuild themselves thus have "their own endogenous dynamics, continually running through their cycles whether perturbed from the outside or not" (Pickering, 2010, p. 164).

This is not to say that the breakdown of internal constraints and other energy stores are the organism's only power supply. Nothing, living or otherwise, can perfectly channel free energy into work, and in every exergonic-endogenic cycle some free energy will be lost through its dispersal as heat. The only reason the organism is able to continue this self-production for extended periods is because of its openness to energy and the continual "top-up" of its reserves from the environment in the form of food or sunlight. Nor is it to say that organismal behavior cannot be influenced by external events. But while these influences shape what an organism does, they are not why it does anything at all. Deprived of either energy or other external promptings, an organism will still be active—right up until the point it loses the ability to channel the *thermodynamic* (!) free energy stored in its own body into work and dies.

The reason why an organism acts is the reason that it embodies in its energetic resources and the precarious constraints that channel this energy into the work on which its existence depends. An organism is genuinely the initiator of its movements in a way in which a machine is not, and thus, if an organism is a reason, then this reason is itself a cause.

12 Wrapping Up

It is only by appealing to both thermodynamic and organizational properties that we can describe the precarious dependence of an organism's existence on its own activity, in virtue of which it constitutes a self-constraining, self-producing system. In doing so, we see why, unlike inorganic structures that can be expressed by universal laws, organic structures as Merleau-Ponty put it, "are understood only by a norm," (1963/1942, p. 148). While a precarious constraint organization will mandate certain interactions as both a consequence and as a precondition for so long as it persists, the organism can also abandon a failing organization in favor of a new one in service of the ultimate norm of self-production by whatever means necessary. Unlike a law, this norm too may genuinely fail, but when it does, the entire existential world of the organism collapses with it.

I have explored this account of autonomy and autopoiesis in terms of constraint closure mainly at the level of simple organisms. Yet, the beauty of the notion of a precarious constraint is that it is at once thermodynamically restrictive enough to avoid attributing normativity and intentionality to any stable mechanism, while being general enough that it could, in principle, be instantiated at all different scales of biological organization.

Still, this may not be enough for you to be willing to consider an organism a rational agent, or even an intentional agent at all. You might take this to require epistemic capacities, like the ability to detach what appears to you from what is and entertain the possibility you might be wrong. You might see it as requiring not only regulation in response to external disruption but the ability to proactively anticipate such disruptions and adjust in response, or to explore the consequences of an action in offline simulation prior to executing it. Or you might require the intersubjective capacity to view others as purposive, intentional agents such as yourself, and to coordinate your behavior with respect to *their* norms.

Such capacities are useful markers with which to judge the conceptual or inferential abilities of a living system. Perhaps they might provide a means to track the emergence of self-consciousness and justify the attribution of it to some animals and not to others. They might provide a means to distinguish between merely acting for a reason and *knowing* that you are acting for a reason. But to propose these as further, necessary, requirements to be an agent is just to say that for a movement to be caused by a reason does not suffice for it to be the act of an agent.

Thus, Hurley (2003), unlike Korsgaard (2018a), argues for the treatment of animals as rational, intentional agents, rejecting the idea of tying practical rationality to either conceptual and inferential capacities, or to something like conscious awareness—neither of which accounts for the *origin* of the normative standards that they are supposed to make us uniquely capable of acting in accordance with. As Hurley argues, the flexibility of behavior that is typically taken as an indicator of conceptual capacity comes by degrees. It does not even require a nervous system. The humble bacterium is capable of adapting its behavioral strategies in pursuit of the ultimate norm of continued self-production—for instance by changing from an organization that must seek glucose to one that must seek lactose when the former set of norms lose viability. Groups of bacteria may even share strategies among themselves, in the form of parcels of DNA-based constraints, called plasmids, that enable the production of other constraints that have proved beneficial in their own case.

This is not to say that bacteria are capable of conceptualization and reasoning or intersubjective and social coordination in the same manner that we are. Only to say that the idea that a transition from “as if” to genuine intentionality might emerge from a gradual complexification of non-intentional activities, as Friston et al. (2020) proposed, will always struggle in identifying a definitive threshold to say *this* is where purposes come into the world. The bioenactive approach offers a much more promising strategy, in reversing the direction of explanation to describe how these more advanced capacities might progressively develop from the adaption and evolution of creatures that are already intentional agents.

In her shared circuits model, Hurley (2008) puts forward an initial proposal for how this might work, describing how a layered feedback control hierarchy could provide a subpersonal mechanism for the capacities of "imitation, deliberation and mindreading." Beginning with the fundamental capacity of online motor control, Hurley explores how the development of a capacity to improve this via offline simulation, in order to predict the potential consequences of one's own actions, enables the possibility to simulate potential actions of others and infer the hidden causes behind these, "thereby enabling strategic social deliberation" (p. 1).

While predictive processing may have initially been proposed as an account of how we infer the distal structure of our environment, its basic requirements for the minimization of prediction error at different levels of detachment and spatiotemporal grain, are the same as for Hurley's account. Reframed in the same enactive terms as the shared circuits model, PP could provide an implementation mechanism that the enactivist can use to explain how more advanced socio-cognitive capacities could emerge out of the more basic capacities of sensorimotor control.

Yet, the coordination of movements to maintain inputs at a stable, predictable state is not, in itself, a *need*. Hurley points to the requirement for a teleological context to ground normative attributions to a control hierarchy but does not supply an account of this teleology. Similarly, nonreconstructive versions of predictive processing have typically gone only halfway toward an enactive account—arguing that the predictive brain is directed toward something other than accurate reconstruction but offering no justification for interpreting the tendency for some neurons to become correlated with others in terms of a purpose that the brain is trying to fulfil. Without this, we have no basis for attributing function or purpose to a "predictive" neural architecture, no sense in which it can genuinely fail, no means to take its dynamics as anything more than the externally determined movements of a law-governed machine, and no reason to consider it any more of a predictive model than one pendulum coming into sync with another.

The free energy framework, and particularly its principle of free energy minimization, initially appeared as a solution for this problem. Said principle, Friston (2012) claimed, could "unify all adaptive autopoietic and self-organizing behavior under one simple imperative; avoid surprises and you will last longer" (p. 2). As free energy is analogous to prediction error, so predictive processing looked nicely placed to provide a mechanism for how organisms achieve this "existential imperative" and to describe how abilities like offline simulation and counterfactual reasoning might be grounded, as

Ramstead et al. (2018, p. 33) put it in “the ‘intentionality’ or ‘aboutness’ of living systems—that is, the directedness of the organism towards a meaningful world of significance and valence,” which, they argue, “emerges as a natural consequence of embedded adaptive systems that satisfy the constraints of the free energy formulation.”

Unraveling this claim was a long, frustrating, and ultimately disappointing experience. What lies beneath the FEP’s twisted thicket of mathematical terminology is just a formulation of survival as stability in the face of perturbation and the insight that such stability can be formally redescribed in inferential terms. An equivalent analysis of survival was formulated by W. R. Ashby half a century earlier who, like Maturana subsequently, saw it as precisely the means to *eliminate* all “metaphysical complications” of purpose, intentionality, teleology, or function from biological explanation. As Ashby argued, any stabilizing thing, whether a computer or a pendulum, can be interpreted as “rejecting” unstable states to “seek out” a stable equilibrium. If intelligence, agency, and intentionality reduce to nothing more than prediction-error-based control, and if such control is attributed to any stable system, then they are either everywhere or nowhere at all.

Pitched as an “existential imperative,” it turned out that the requirement of stability had even worse problems than its generalizability. While the requirement might be softened in various ways to make it more trivial, living systems are precisely those things that are most likely to violate it. Modeling an organism as a free energy minimizer may work well for describing some of its behaviors over some timescales, but the properties that define this model do not define the identity of the living system itself, which is liable to change in ways that cannot be predetermined by its organization alone. This is not just a problem for the FEP, however, but for any attempt to define living systems in terms of the invariant logic of a machine.

The prevailing formulation of autonomy within the enactive approach is more amenable to the kind of open-ended change that we observe in the development and evolution of living systems. Yet, for bioenactivism to succeed, this definition of autonomy needs to include a robust enough notion of self-production, such that we can follow the organism through these organizational changes, account for what renders its existence fundamentally different from that of a machine, and license a realist attribution of agency and intentionality to organisms alone. Process closure, I argued, fails in these respects.

This does not mean that the bioenactive project fails too. As I have described, Moreno, Mossio and Montévil's work in constraint closure succeeds on all three counts, providing a naturalistic grounding for the application of normative notions such as purpose, intentionality, goals, or function to living systems alone. Moreover, their account of regulation in terms of second-order constraints provides the opening for extending these intentional attributions to extra-metabolic systems, like the nervous system, in terms of the role these play in coordinating the transition between different constraint-closed organizations in order to preserve an ongoing process of production.

Insofar as the organism, and its brain in turn, is trying to achieve the continuation of a process by whatever means necessary—rather than the stability of some particular set of states or relations—so its function will not be reducible to free energy minimization. Yet, as I acknowledged, the FEP is still useful in building approximate models of a system. A wide range of organismal behaviors are, after all, homeostatic. If this homeostasis and stability are not *necessary* for survival, then why would they be so common?

A potential answer to this may be found not in enactivism, but rather in the epistemic anxiety of Hohwy and Helmholtz. Insofar as we cannot determine in advance whether some new interaction or organization will support the continuation of constraint production, so, even as enactivists, we face the skeptical challenge of attempting to coordinate ourselves with respect to unknown factors. To do so, living systems need to make a guess at how things around them might unfold. The simplest one is to presume that they will carry on just as before—and that if a particular organization has proved viable in the past it will continue to be so in the future. On such an assumption, it makes sense to attempt to keep the state of ourselves and our environment within the stable bounds that this particular organization requires for as long as possible, before taking the risky leap into a new form.

So, survival may not be equivalent to the minimization of surprisal, but, when in doubt, avoiding surprisal may still help secure it. Just as this free energy approximation can prove useful to scientists seeking to model living systems, so it may be a useful way for organisms to approximately model themselves, for the purposes of predictive control.

That evolution of this into *hierarchical* predictive control may lead to all sorts of wonderful things like the ability to predict others, to entertain different possible expectations about what will happen, to try out inconsequential predictions, make non-fatal mistakes, and to live through this discovery that we can get things wrong. Still, what matters is not the complexity of the control hierarchy required for these capacities, nor the variety of perturbations it can respond to, nor how many different stable states it might alternate between. What matters is the precarious status of the intrinsically unstable organism that both realizes it and depends on its successful operation.

We could copy this hierarchical control structure and grant it the physical “embodiment” of effectors and sensors hooked up to the sort of inputs that matter to us. Yet insofar as this system is built out of parts that have nothing to lose by its failure, any interpretation of it as trying or failing will only be a projection of our goals onto the movements of an utterly indifferent machine.

Appendix: What's the Use of a Concrete Blanket?

Over the past year, a growing number of critics of the FEP have begun to object to the ease with which papers by Friston and coauthors slide between the standard heuristic formulation of a Markov blanket and a stronger ontological one (Menary & Gillett, 2020; Bruineberg et al., 2021; Raja et al., 2021).

As described in section 5.2, for instance, Allen and Friston (2018) move from the uncontroversial statement that "the boundary (e.g., between internal and external states of the system) can be described as a Markov blanket" to the description of a Markov blanket as a real boundary that the system must actively conserve, and upon which its very existence causally depends. Similarly, Ramstead et al. (2018) switch freely back and forth between the Markov blanket as way of describing some statistical relationship in the language of a Bayesian network, versus its being the thing in the world that produces the very conditional independence that the model then describes:

> Markov blankets establish a conditional independence between internal and external states that renders the inside open to the outside, but only in a conditional sense (i.e., the internal states only "see" the external states through the "veil" of the Markov blanket . . . With these conditional independencies in place, we now have a well-defined (statistical) separation between the internal and external states of any system. A Markov blanket can be thought of as the surface of a cell, the states of our sensory epithelia, or carefully chosen nodes of the World Wide Web surrounding a particular province. (p. 4)

The location of a Markov blanket, they go on to say, provides the basis for "a fully generalizable *ontology* for biological systems" (p. 5) [my emphasis]. This realist interpretation is not a part of the original concept of a Markov blanket, and neither does it just follow inevitably from the mathematical core of the FEP.

As Bruineberg et al. (2021) point out, “Metaphysical consequences require metaphysical premises, and cannot simply be read off the formal model.” At first appearances, then, the reification of Markov blankets looks like a classic case of confusing properties of the model for properties of the target system. Such missteps, Andrews (2021) argues, are endemic in the FEP literature where terms like “entropy” or “energy,” appropriated from thermodynamic systems to describe analogous statistical properties, are then misinterpreted as retaining implications that follow only in the limited context of describing constrained flows of matter and energy in concrete physical systems.

Realism about Markov blankets might thus be taken as just another instance of this “fallacy of misplaced concreteness” (Whitehead, 1925). This is an ever-present danger with scientific models, where utility often comes apart from representational fidelity (Morgan & Morrison, 1999; Potochnik, 2017). Still, if realism about Markov blankets is indeed a fallacy, this would not undermine the FEP’s legitimacy outright. Andrews (2021), for instance, suggests shedding this extraneous pretention to describing an objective feature of all living systems, in order to separate out the mathematical core of the FEP as a purely formal model structure, from both conceptual and empirical questions regarding its applicability to any particular target system.

Had advocates of the FEP stuck to this more abstemious interpretation of its key constructs, then their work would have likely generated far less confusion and controversy. It would also probably have garnered much less attention for its authors. Unfortunately, as we have seen, this distinction between the pure formalism and the various philosophical theories and models that have been constructed upon it, has rarely been respected within the FEP literature. Claims that the FEP might provide a first principle of living systems, one that could subsume autopoiesis and autonomy, fall squarely within the purview of philosophy and theoretical biology—not mathematics. In critically analyzing the FEP as a theory of life, it is specifically these extraneous philosophical claims that I’m concerned with. As regards Markov blankets, said theory of life requires that we can indeed consider these to be a real entity that is possessed by living systems, and living systems alone, which has causal influence to the effect of preserving and distinguishing them from their surroundings.

In the interests of steel-manning this view, it’s worth pointing out that while proponents of ontological MBs have not provided the argumentation to support this realist position, neither have the FEP’s critics shown that such realism is incorrect—only that the justification for it is currently lacking. The fact that (as discussed in chapter 5) those Markov blankets that appear in our model are sensitive to a host of initial choices—such as which variables to use and the coarseness of grain with which these were individuated—does not entail that there is not genuinely something like a set of "real Markov blankets" in the actual structure of the world. Nor does it preclude the possibility that our partial models and often incorrect models may sometimes allow us to accurately identify them. It may be a platitude of the modeling literature that "all models are wrong," but it is presumably also true that all useful models must be getting at least something right. Some features of our maps must really be features of the territory, or they will not be maps at all.

So, if Friston and colleagues have not supplied the metaphysical premises required to support their metaphysical conclusions, can we derive these on their behalf?

A.1 God's Great Causal Graph

What would a causal modeler have to say about the reality of Markov blankets? Well, firstly, they would likely point out that the fact that every causal graph has them is simply a consequence of the presuppositions that direct such a graph's construction. For our purposes, the relevant ones are the following: (1) **Decomposability**: the system can be broken down into a set of discrete parts and the connections between them, and (2) interactions between these parts respect **the Markov condition**, such that the state of each is independent of the state of its non-descendants, conditioned upon its parents. This second criteria ensures the factorizability of our graph in terms of Markov blankets and becomes the *causal* Markov condition when said graphs are interpreted as causal models.

So, the issue of whether Markov blankets are merely a "statistical device" or a "necessary attribute" of "the system itself" comes down to whether these are simply useful falsehoods to help us approximate the behavior of said system, or whether they accurately describe its structure and interactions between its parts.

Among the developers of programmatic causal discovery, these axioms of construction have primarily been defended on a pragmatic basis, not as either analytic truths about our concept of causation or principles reflective of some fundamental law of nature (Spirtes et al., 2000; Pearl, 2000; Stafford, 2005; Weslake, 2006). Objections to the metaphysical implications underlying such models tend to be overlooked, if not met with outright hostility (Glymour, 2010). Thus, we find Glymour (1999) defending the causal Markov condition's legitimacy, not via what he demeans as "Socratic analysis," but rather by means of a list of cases in which methods premised on it have been able to predict the result of interventions from observational data alone. As he puts it, "The essential issue in scientific discovery is the right representation for reliable, efficient search, not the metaphysical disputes upon which philosophy of science is fixated" (p. 64).

Still, while the causal modelers may not be overly concerned, the position that these axioms of construction might be more than pragmatic—the proposal that they describe a necessary feature of reality—is one that predates the FEP. Decomposability, the claim that reality is built up out of individual building blocks, has been a (surprisingly) resilient cornerstone in the history of science and philosophy, from the atoms of Leucippus to the individuals of classical logic. Its resilience is surprising because the status of this long-standing doctrine as a description of fundamental reality has become increasingly threatened by developments, such as quantum field theory, that suggest that such apparently "elementary" particles may not constitute basic reality after Hobson (2013). As Bickhard (2015) puts it, "According to our best physics, there are no particles and what are called particles in contemporary parlance are quantized excitations in quantum fields" (p. 24).

The idea that the Markov condition is no mere modeling tool, but rather a constraint that actual causality must satisfy, does not have quite so long a heritage. Still, it can still be traced back through philosophical discussion of the Markov condition's ancestor in Hans Reichenbach's (1956) analogous "principle of the common cause,"[1] proposed not as an optional heuristic but rather a necessary requirement governing the relationship between correlation and causation that Reichenback attempted to place on a par with the second law of thermodynamics.

In an attempt to give more metaphysical bite to Spirtes, Glymour and Schienes' models, Hartry Field nicely expresses what reality might look like if one accepted these two tenants, suggesting that:

> Intuitively, it seems (barring quantum non-locality and the like) that one should be able to think of the physical universe as a causal system with a node for each space-time point, with the value of the node expressing the totality of the values

> of physical quantities at that point; the light cone structure gives the dependence relations. (2003, p. 447)

I will call the position that causal graphs accurately capture the structure of reality causal graph realism.[2] For Field, the fundamental nodes of reality are microphysical space-time points, or point-sized particulars, that cannot be further decomposed, a position defended more extensively by David Lewis (1994) as "Humean supervenience." But being a realist about causal graphs does not necessarily require the commitment to microphysical reduction. As part of his defense of the reality of non-reducible macrophysical phenomena, Papineau (1992, 2022) takes causal graphs to provide an accurate metaphysical picture, but one that concerns relations between generic states of affairs—for example, between the prevalence of smoking and prevalence of early mortality in a population—and which cannot be reduced to microphysical dynamics (where causal relations look to be absent altogether).

The causal graph realist need not be troubled by the fact that many of our causal graphs will be wrong, or that the inevitable simplifications made in the construction of such models invariably divorce them from the real structure of the world. As Spohn (2001) argues, we can still conceive, in principle, of a graph constructed with an "all-encompassing frame" that would capture *all* of the correctly individuated units and their direct relations (and so, ultimately, all the "ground-truth" Markov blankets). Such a graph, he argues, would capture all there is to say about causation. The correctness of a particular simple model, for the causal graph realist, is simply a matter of its similarity to this all-encompassing graph.

Causal graph realism may (but need not be) combined with support for the *statistical reduction of causation*,[3] to claim that the correct causal graph contains nothing more than all the statistical relationships between our various nodes.[4] This reduction is developed by Spohn (2001), who explicitly extends Spirtes, Glymour and Schienes' work to argue that "Bayesian nets are all there is to causal dependence." While the idea is essentially Humean, similar reductions have been developed by Reichenbach (1956), Good (1959), Suppes (1970), and Papineau (1992) (for reviews, see Salmon, 1980; Weslake, 2006). On such accounts, the gap between correlation and causation is not due to their being different things entirely, but rather due to the incompleteness of our correlational information rendering it insufficient to determine a single causal graph.

These accounts differ, however, on whether they consider the reduction to be conceptual (Spohn, 2001; Suppes, 1970) or metaphysical (Papineau, 1992, 1993; Field, 2003). Where for Suppes, probabilistic/causal relationships are relative to a scientific model, the metaphysical reductionist is committed to there being a true set of probabilistic relations that make up some objective causal graph.[5] On this view, the all-embracive Bayes net does not just describe how we conceptualize causality, but correctly describes the structure of reality itself—the axioms of causal modeling are neither contingent facts about the relationship between causation and probability nor analytically true in virtue of how we conceptualize causality, but simply describe what causality actually is.

Let's call the combination of the strongest version of these two positions "probabilistic graph realism." There are a host of difficulties in holding such a view. With regard to the probabilistic reduction of causation, the most obvious is the need for a robust account of objective probabilities. It's also interesting to note that while Pearl (2000) or Glymour (2010) avoid taking positions on the metaphysics of causation, they explicitly reject a probabilistic reduction as being sufficient even for the methodology of causal investigation, let alone an ontology of causal relationships. Instead, Pearl (2001) advocates an "interventionist" account (Woodward, 2005), arguing that disambiguating unique causal structures requires going beyond mere statistical relationships to include the effect of targeted interventions to fix the value of a particular variable. In more concrete terms, the kind of disambiguation of causal relationships that we achieve by moving the needle on a barometer and seeing if it causes a thunderstorm.

Other difficulties include deriving the temporal asymmetry required for a causal graph (and crucial in the FEP's division of the Markov blanket into parent/child nodes). This cannot be simply delegated to fundamental physical laws, which are time-symmetric, and it is controversial whether it can be derived either from within purely statistical asymmetries or from some other external source, as in Reichenbach's enlistment of the second law of thermodynamics (Price, 1993). Further, if we take the units of Spohn's "all-embracive frame" to be microphysical (for instance the local spacetime points of Field [2003] and gr [1994], we not only lose temporal asymmetry but also decomposability and Markovian behavior, which are upended by quantum entanglement and indeterminacy (Glymour, 2006; Arntzenius, 1992; Cartwright & Jones 1991; van Fraassen, 1980, 1982).

Nonetheless, I think that interpreting the realist manner in which Friston talks about Markov blankets to be a product of an implicit commitment to the metaphysics of probabilistic graph realism is a more charitable, and more plausible, interpretation than taking him to be merely mistaking the constraints of a particular modeling framework for real features of modeled systems.

This implicit metaphysics makes sense of more than just Markov blanket realism. It also bears on the discussion in section 3.4 about the status of the generative model and claims that the interaction between an organism and its environment is not only described by, but literally embodies, this joint distribution. A statistical graph, recall, represents a joint distribution, the generative model, plus a (typically temporal) ordering. To say that the structure of reality is such a statistical graph is to say that it actually has the properties of this generative model. So, to say that systems literally are generative models, and that they really have Markov blankets as their parts, is not to say that these systems are representations of anything else, or that they are tools being used to some epistemic end. It is to say that the system itself has the same kind of properties as the partial graphical models that we build of it. Under this particular metaphysical view, the gap between the model of a scientist and the real system itself would be merely in degree of detail and not a difference of kind.

A.2 Naturalized Mathematical Realism

Still, even if this delivers objectivity to claims about Markov blankets as a description of the true causal graph, does this confer independent existence upon the Markov blanket itself? Majid Beni (2021) argues not, pointing out that Markov blankets are still a mathematical object. As such, he points out, to treat them as something that exists in its own right, rather than as an accurate description of a really existing *physical* object, must either be a category error, or the result of a commitment to realism about mathematical objects.[6] A similar point is raised by Bruineberg et al. (2022), who argue for a distinction between treating the Markov blanket as "literally" an entity in the world and seeing them as "realist" descriptions of some other feature that does have worldy existence.

A commitment to the FEP as providing a "formal ontology" as advanced by (Ramstead et al, 2021), who propose its use as "a mathematical formalism to answer the questions traditionally posed by metaphysics; i.e., what does it mean to be a thing that exists, what is existence, etc.," does imply a commitment to some form of mathematical realism. However, both Beni (2021), and Menary and Gillett (2020) in a similar criticism, interpret this realism as of the Platonic form. Platonism may be the best-known mathematical realist position, but its defining feature is the treatment of mathematical objects as transcendent entities existing outside of space and time. That might fit with understanding the FEP as a purely formal model that nevertheless says something true, as Andrews (2021) suggests we take it. But this cannot be the kind of reality Friston takes a Markov blanket to have if it is also intended to be a property or thing that particular systems really have that can enter into the kind of causal relations that preserve that system's existence.

As applied to living systems, the FEP seems to require the Markov blanket to be a simultaneously physical and mathematical entity. As such, I suggest this application of the FEP would be better paired with the kind of naturalized mathematical realism developed by Penelope Maddy (1990, 1997), under which there are physical objects that instantiate the properties of a mathematical object—specifically, for Maddy, the property of being a mathematical set. If the world is indeed structured like a statistical graph, respecting the Markov condition, then for any system, there will be a set of objects that have the property of rendering the state of some further set of things probabilistically independent of everything else. Naturalized mathematical realism would allow Friston to describe this set as a physical object that literally has the mathematical property of being a Markov blanket.

A.3 Absolute Units

We now have the elements that seem to be behind the principled commitment to Markov blankets: causal graph realism, the probabilistic reduction of causation, and naturalized mathematical realism. Let's say we adopt all those. There really are physical sets that render a particular thing independent of everything else. Where exactly are these "Markov blankets" then? Are they in the room with us right now?

The answer to this depends on a final missing detail in Friston's (supposed) causal graph realism: If reality is indeed divisible into discrete individual units, what are those units? We've seen that, for Friston, a Markov blanket is at once a mathematical object and something that exists in the physical world. As such, each node in the network that composes it must itself be the state of some particular physical thing, and the graph will be constituted by statistical regularities holding between the changing states of these particulars. This is another reason why his use of the tools of causal graphical modeling may look strange to those familiar with the causal discovery work of Pearl, or of Sprites, Glymour, and Schienes, where causal-statistical relationships are modeled as holding not between token objects but between general events, properties, or states of affairs (Hausman, 2005). Here, each node in the graph is not a concrete particular but a type that may be instantiated by many different particulars.

Friston, in contrast, takes nodes of the graph to be the token states of concrete parts of the physical world. Unlike Lewis and Field, however, he does not commit himself to the ultimate causal graph being microphysical. In Friston (2013), the relevant units are indeed the electrochemical and kinetic states of particles in a primordial soup, but in Hipólito et al. (2021), they are the states of synapses and ion channels, and in the proposed models of societies and ecosystems, each node could be the state of an individual organism (Kirchhoff et al., 2018).

The idea here, as Ramstead et al. (2019) explain, is that what appears as the single node of one Markov-blanketed system is itself a Markov-blanketed complex that may be further decomposed into sensory, active, and internal components. We might call this *multiscale probabilistic graph realism*: the position that at whatever scale we consider a system, it will be separable into distinct components whose interactions respect the Markov condition. Which scale we pick, Ramstead et al. argue, will be relative to our interests, but the divisibility into Markov-blanketed subsystems is not. Once all the components at that scale are included (even if this is not actually possible in some particular simple model), then there will be a set of *real* Markov blankets for that scale. As Sims (2021) suggests, such a multiscale formalism may actually be better placed to capture biological and physiological individuality, in which individual organisms—a somatic cell or a symbiotic organism—cooperate with others to form a higher-level individual, while still preserving their own individuality at a lower level of analysis.

While Friston (2019a) suggests this may apply "ad infinitum," that causal graph structure may persist "all the way down" (p. 7), this commitment is not integral to the FEP, and later in the monograph, he expresses "metaphysical agnosticism" on the issue (p. 124). We thus have the option to take Papineau's (1992) approach and reject the requirement for macroscopic phenomena to be reducible to microscopic interactions, sidestepping the spanner that quantum mechanics throws into the workings of a microphysical causal graph, while still maintaining the reality of individual units, and the Markov blankets they compose, as an emergent macrophysical phenomenon.

A.4 A Second Stability Requirement

Our intitial problem with Markov blankets was that they might be drawn anywhere, depending on how we construct our causal model. Once we move to the all-encompassing graph, the constitution of Markov blankets may not be arbitrary, but they are nonetheless pervasive. If we take causality to be a local phenomenon, as presumed in Friston (2013), such that only contiguous elements can interact directly, and if we treat causality as reducible to statistical relationships between these elements, then for *any* element that we pick, the state of its immediate surroundings will render it conditionally independent of everything else in the world. If every organism is surrounded by a real Markov blanket, this is not the result of its own efforts, but merely a general result of a particular account of the metaphysics of causality.

Friston (2013, 2019) does not seem to take the existence of a Markov blanket to be quite as trivial as I have outlined. A candle flame cannot possess a Markov blanket, he argues, because in contrast to a cell membrane, "any pattern of molecular interactions is destroyed almost instantaneously by the flux of gas molecules from its surface" (2013, p. 2).

> In other words, it [The Markov Blanket formalism] does not easily accommodate the fact that the particles that constitute a Markov blanket can, over time, wander away or, indeed, be exchanged or renewed. The canonical example here would be the blanket states of a candle flame, whose constituent particles (i.e., molecules of gas) are in constant flux. (2019, p. 50)

This, Friston (2013) claims, is contrasted with "the physical configuration and dynamical states that constitute the Markov blanket of an organism—or organelle—change slowly in relation to the external and internal states it separates" (p. 10). So, what he seems to require here is that the connections between various parts of a system are stable throughout the changes in states of those parts, such that the particular set of elements that make up the Markov blanket is conserved over time. Indeed, if the way we make sense of objective probabilities is in terms of something like long-run frequencies or propensities, then the connections between units *must* be more stable than the changes in the state of those units in order for there to be a statistical relationship between them, and so for there to be a graph divisible into Markov blankets at all.

So, the systems of interest to the FEP are not supposed to be distinguished by the property of having Markov blankets per se, but rather by having the same particular Markov blanket with constituents that are stable[7] over some duration. This helps explain the conflation Friston's critics accuse him of, between describing the Markov blanket as a "physical boundary" and "statistical partition" (Bruineberg et al., 2022). On the putative metaphysics I'm ascribing to Friston, a Markov blanket ends up being both. If reality literally is a statistical graph, and if causal interactions are exclusively local, then a stable Markov blanket is both a statistical partition and one realized by a fixed set of discrete elements that surround the system of interest—in other words, a boundary.

We can thus think of the requirements for a system being a free energy minimizer in terms of two types of stability. Firstly, the stability of the typical state of its parts, derived from the analysis of survival as surprisal minimization, and secondly, the stability of statistical-causal dependencies between these parts that allows us to identify a particular Markov blanket as persisting over time.

These then, are the steps required to claim that Markov blankets are real things. All that positing a real Markov blanket involves is claiming that a system decomposes into independent units (accomplished by some means prior to the FEP, by methods the framework itself does not specify) and arguing that if we accept objective statistical dependencies as an adequate reduction of causal relationships, along with the principle of locality, then the immediate surroundings of any one of those units will literally have the Markov property of inducing a conditional independence between what is inside this boundary and what is outside of it. To claim that Markov blankets are real things is to make a general claim about the structure of the causal universe. In doing so, the FEP has neither identified a new and interesting entity in the world nor discovered a principled way of carving the world into things. This latter task depends on the necessarily prior task of telling us what the absolute units of the ultimate graph are.

Notes

Chapter 1

1. Thompson chooses instead to modify the notion of autonomy to incorporate adaptivity within it. Nonetheless, there is agreement that it is at least not derivable from the example of autopoiesis alone.

Chapter 2

1. It's interesting to note here that while the above accounts take "inferential" and "Bayesian" to be synonymous, Helmholtz's concept of inference was associationist, not Bayesian. (Westheimer, 2008). Like Kant, Helmholtz sought to combine a recognition of the inescapable contribution of the cognizer in constructing the objects of experience with an "objectively valid" basis for these constructive or inferential processes, such that we could be said to have knowledge of the objects they present to us. Yet, while Helmholtz critiqued Kant's own attempt to identify these a priori "laws of thought," particularly in terms of the contingency of Euclidean space, he struggled to provide an adequate alternative that might confer validity on our inferential processes (see Hatfield, 1984, 1990). It would thus be more accurate to view Bayesian theories of perception as an advancement on Helmholtz's account, where Bayes's rule is postulated as providing this objectively valid law of cognition. The question for the reconstructivist interpretation of PP is how much accordance with such a law guarantees regarding the convergence of our beliefs on an accurate representation of some mind-independent reality.

Chapter 3

1. For critical reviews of instrumentalism versus realism in the FEP, see Bruineberg et al. (2022), Andrews (2021), van Es (2021), and van Es and Hipólito (2020).

2. Also known as variational Bayes, or ensemble learning.

3. As discussed with regard to PP, these causes of changes in the sensory stream involve the brain's own actions. For now, we're ignoring the possibility for the

system to intervene on the process that generates its evidence, to get through the basics of variational inference as an approximation technique.

4. Also called the recognition density, or variational density. Density is simply the term for a function that generates a continuous probability distribution.

5. Not to be confused with thermodynamic free energy!

6. The "< > q" symbols denote that we're taking the weighted average for the probability of o with respect to the recognition density q(H) called the "expectation" or "expected value."

7. A neologism coined to distinguish the notion of unlikeliness, from the personal-level notion of surprise.

8. by VisitScotland officials at the least.

9. Indulge me in imagining yourself somewhat geographically challenged.

10. The remarks on Hohwy's view here pertain to the "Helmholtzian" treatment given in his 2013 book *The Predictive Mind* and contemporaneous papers. It should be noted that in more recent works, he appears to move away from this approach to an understanding of PP more compatible with the current treatment.

11. While in the PP literature, it is common to divide the prediction error minimizing process into "active inference" as opposed to "perceptual inference" in this way, in the FEP the term "active inference" is used to refer to the unifying story that combined both perception and action together in the minimization of variational free energy.

12. In chapter 5, I'll describe how we might get some constraints on possible patterns of influence out of a joint probability distribution. As we'll see, this determination of a direction of influence always requires further assumptions or information not contained in the generative model itself.

13. Ergodicity is a stronger requirement than the stability of statistical properties over some duration, and neither is trivial, but we can put this aside until chapter 4,

14. There are actually two divergences making up misrepresentation in Hohwy and Kiefer's argument, as they distinguish between a generative model encoded in the brain and the true statistics of the generative process. If we follow the arguments of the previous section, however, then there is only the true generative model and only one relevant divergence term between this and the recognition model.

Chapter 4

1. The entropy and variance of a distribution, while related, are different things. However, if (as mentioned in chapter 3) we are constraining our probability

distributions to single peak Gaussians, then the only way to change the entropy of a distribution is to change its variance.

2. More recent work has suggested even the requirement of steady state can be abandoned (Friston et al., 2023; Ramstead et al., 2023). In chapter 9, I will argue that abandoning this requirement would render the free energy principle vacuous.

3. This claim is sometimes explicitly formulated in terms of "generalized coordinates of motion" where the state space includes not just states like "position" but also derivatives of these states, such as velocity or acceleration. All the following arguments regarding the role of steady states in the FEP apply equally where these "states" incorporate stable rates of change, rates of change of rates of change, and so on and so forth.

4. As Ward (2016) argues, in Susan Hurley's (2001) synthesis of Power's perceptual control theory with the motor theories of perception we have an even closer forerunner of predictive processing.

5. While Friston et al. (2020) primarily propose this as a gradualistic criterion for the emergence of consciousness when taken as a metric of autonomy or animacy, it may serve as a defense against hylozoistic consequences too.

6. In chapter 9, I will discuss more recent developments under which this commitment to steady states has been relaxed also. Such a weakened interpretation of the free energy principle may grant Friston's framework the flexibility it needs to better model living systems, but only at the cost of admitting that the principle itself places no meaningful constraints on their potential dynamics.

Chapter 5

1. The graph theorist would say "vertices" and "edges," but I will stick with the more familiar vocabulary of nodes and connections.

2. It is not immediately obvious why the parents of children are parceled up as "active states" here, given that they themselves may not be influenced by internal states at all.

Taking Stock

1. This was originally secured via the stronger claiming that systems are ergodic, but while ergodicity entails stationarity, it also entails that a system will eventually explore every state that it is possible for it to be in—such that the probability distribution describing the states of a particular system over time converges to the distribution that describes the states of an ensemble of that type of system at a single moment in time. This is a stronger claim than needed, and one inappropriate for living systems (see Palacios and Colombo, 2021).

Chapter 6

1. This ordering already introduces more than the existential dyad strictly buys us, for a single ordering into ancestors and descendants will not always be supplied by statistical relationships alone.

2. Co-parents of children, though a part of the Markov blanket, are typically ignored—presumably due to the difficulty of fitting them into the category of either sensory or active states.

3. The parents of children are more difficult to incorporate in the FEP framework and are typically ignored.

4. Crucially, like the Markov blanket. however, these labels will still depend on the prior selection of some set of states as “internal” ones. In the ESIA cycle, internal and external states are symmetrical, and there is nothing to prevent us swapping their labels, along with those of action and sensory states.

5. This approach to living systems as cycles upon which an observer may impose an organism/environment distinction that does not originate from the cycle itself, is a direction continued in Humberto Maturana’s pre-enactivist development of autopoeisis theory. While autopoiesis theory comprises the genesis of a great deal of bioenactivism’s conceptual repertoire (not least with regard to autopoiesis itself), the two should not be confused, and indeed take radically opposed views on the status of life. Crudely, while the former retains this Ashbyian view of the continuity between life and non-life, the latter originates in Varela’s later phenomenological, organicist, or existential turn, which grants organisms special status as “natural purposes” capable of “bringing forth a world of meaning.” More on this, in section 7.4.

Chapter 7

1. As I’ll discuss in chapter 9, this focus on closure of, or among, processes exclusively has helped prevent an adequate formalization of autonomy that in Mossio and Moreno’s (2015) terms, requires closure between two regimes of causation involving both change and invariance.

2. The history of these two terms would require work beyond the scope of my argument here—particularly given the ambiguity over attribution that Maturana and Varela’s coauthorship entails. It seems to me that their work does not consistently follow the terminological distinction Bich and Arnellos suggest, and moreover I’m not convinced that “organizational” is the best choice to distinguish the closure of a self-producing system from that of a merely operationally closed one.

Chapter 8

1. As discussed in the section on the metaphysics of Markov blankets, I am allowing the conflation of causal relationships with statistical ones here, based on the argument that in some complete and comprehensive causal network, the former will reduce to the latter.

2. As Proksch (2021) notes, this requirement of stability is equally difficult to sustain in cases where Markov blankets are supposed to serve as the boundaries of social networks, given that the turnover of members is a feature of most social organizations.

Chapter 9

1. The authors also emphasize another sense in which the free energy principle can be rendered "path based" in by being formalized in terms of "generalized coordinates of motion" which involves incorporating derivatives of positional states, such as velocity or acceleration. Formulated in such terms, the stationary probability distribution sought by the FEP might describe not a steady "state" but rather a steady rate of change, or a steady acceleration in a rate of change. Still, steady rates of change—or any steady higher-order derivatives of the same—are no more plausible as a fixed invariant for organic systems. The preceding arguments all apply to an organismal state space that includes such states of motion.

2. For a real system, we would also have to incorporate a coefficient of friction to describe how total energy is dissipated by this resistance, but the trajectories of the pendulum would still be constrained such that at each point the greater the angular displacement, the lower the velocity—and vice versa.

Chapter 10

1. While DiFrisco reserves "substantialism" for the atomistic view, contrasted with a formalist notion of substance, his criticism is targeted at both in terms of their commitment to invariant features. Thus, as he notes, both would count as substantialist in the persistence sense.

2. While it may be less of a distortion, it is still an idealization in that energy transfer via work is never perfectly efficient and energy loss via heat in the course of the operation of a machine can eventually degrade the structure of the machine itself.

3. In her two-part review of the history of the concept of self-organization, *Organisms, machines, and thunderstorms*, Evelyn Fox Keller (2008, 2009) provides a lovely overview of the development of these contrasting approaches to the organism.

Appendix

1. The differences in content between the causal Markov condition and the principle of the common cause are not important for our purposes, but for more on this, see Hausman and Woodward (1999).

2. This shouldn't be confused with a holistic graph realism in which there are no properties of individual particulars, but all properties supervene solely on the relations instantiated in the graph.

3. This statistical reduction of causation is not incompatible with a deterministic account, as Papineau (1989) notes in defending both. Determinism merely involves raising the threshold of statistical relevance for one thing to be a cause of another, such that A only actually causes B if its occurrence raises the probability of B to 1.

4. A causal graph realist doesn't have to adopt this, however. They might maintain that all the statistical facts in the world would not suffice to determine the true causal graph, and that additional facts—regarding say counterfactuals, the effect of interventions, or temporal asymmetry—are required. To defend a probabilistic reduction of causation requires arguing that these facts, insofar as they pertain to causation, may also be reduced to statistical ones.

5. This of course requires a worked-out notion of objective probabilities, but that's a whole other kettle of fish that we don't have time to poach.

6. Such a position would not be Platonist but Pythagorean.

7. This requirement of stability is worryingly vague to be playing such a key role in determining whether a system "exists" or not—but there are only so many rabbits an author can be expected to chase.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50.

Aguado-Velasco, C., & Bretscher, M. S. (1999). Circulation of the plasma membrane in Dictyostelium. *Molecular Biology of the Cell*, *10*(12), 4419–4427.

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459–2482.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), Philosophy and predictive processing (p. 4). MIND Group.

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences*, (3), 204.

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy, 36*(3), 30.

Arntzenius, F. (1992). The common cause principle. *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, 2, 227–237.

Ashby, W. R. (1940). Adaptiveness and equilibrium. *Journal of Mental Science*, *86*, 478–484.

Ashby, W. R. (1948). Design for a brain. *Electronic Engineering*, *20*, 379–383.

Ashby, W. R. (1949). The electronic brain. *Radio Electronics*.

Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall.

Ashby, W. R. (1962). Principles of the self-organizing system. In H. Von Foerster & G. W. Zopf, Jr. (Eds.), *Principles of self-organization: Transactions of the University of Illinois Symposium* (pp. 255–278). Pergamon Press.

Atkins, P. W. (1984). *The second law*. Freeman.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Austin, C. J. (2020). Organisms, activity, and being: On the substance of process ontology. *European Journal for Philosophy of Science*, *10*(2), 1–21.

Badcock, P. B., Friston, K. J., Ramstead, M. J., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: An evolutionary systems theory of the human brain, cognition, and behavior. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(6), 1319–1351.

Bak, P., Tang, C., & Wiesenfeld, K. (1988). Self-organized criticality. *Physical review A*, *38*(1), 364.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: An active inference account of Watt governors. arXiv preprint arXiv:2006.11495.

Barandiaran, X. E. (2017). Autonomy and enactivism: Towards a theory of sensorimotor autonomous agency. *Topoi*, *36*(3), 409–430.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B*, *371*(20160011), 1–17.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive & Affective Neuroscience*, *12*(1), 1–23.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*(1), 217–234.

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Order No. 10016139). [Doctoral dissertation, University College London]. ProQuest Dissertations & Theses Global.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 421–441.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Pačes, J., Burt, A., & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of Sciences*, *101*(14), 4894–4899.

Beni, M. D. (2021). A critical analysis of Markovian monism. *Synthese*, *16*, 1–21.

Bich, L., & Arnellos, A. (2012). Autopoiesis, autonomy, and organizational biology: Critical remarks on “Life after Ashby.” *Cybernetics & Human Knowing, 19*(4), 75–103.

Bickhard, M. H. (2000). Autonomy, function, and representation. *Communication and Cognition—Artificial Intelligence*, *17*(3–4), 111–131.

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology*, *27*, 75–84.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Science*, *17*(12), 656–660.

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A technical critique of some parts of the free energy principle. *Entropy*, *23*(3), 293.

Bitbol, M., & Luisi, P. L. (2004). Autopoiesis with or without cognition: Defining life at its edge. *Journal of the Royal Society, Interface*, *1*(1), 99–107.

Boden, M. A. (1999). Is metabolism necessary? *The British Journal for the Philosophy of Science*, *50*(2), 231–248.

Boden, M. A. (2000). Autopoiesis and life. *Cognitive Science Quarterly*, *1*(1), 115–143.

Boghossian, P. (2014). What is inference? *Philosophical Studies*, *169*(1), 1–18.

Borges, J. L. (1998). On the exactitude of science. *Collected Fictions* (Andrew Hurley, Trans., p. 325). Penguin.

Bourget, D., & Chalmers, D. J. (2021). *Philosophers on philosophy: The 2020 PhilPapers survey*.

Bourgine, P., & Stewart, J. (2004). Autopoiesis and cognition. *Artificial Life*, *10*(3), 327–345.

Bourgine, P., & Varela, F. J. (1992). Towards a practice of autonomous systems. In *Toward a practice of autonomous systems: Proceedings of the first European conference on artificial life* (Vol. 1). xi±xvii. MIT Press.

Brandom, R. (1979). Freedom and constraint by norms. *American Philosophical Quarterly*, *16*(3), 187–196.

Brandom, R. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkte* [Psychology from an empirical standpoint]. Leipzig: Duncker & Humblot.

Brown, H., Friston, K. J., & Bestamnn, S. (2011). Active inference, attention and motor preparation. *Frontiers in Psychology*, *2*, 218. https://doi.org/10.3389/fpsyg.2011.00218

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new Markov blankets. *Behavioral and Brain Sciences*, *45*, 1–63.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6), 2417–2444.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Burrows, M., Rogers, S. M., & Ott, S. R. (2011). Epigenetic remodelling of brain, body and behaviour during phase change in locusts. *Neural Systems & Circuits*, *1*(1), 1–9.

Cao, Y., Mu, H., Guo, J., Liu, H., Zhang, R., Liu, W., . . . & Liu, H. (2020). Metabolic engineering of Escherichia coli for the utilization of ethanol. *Journal of Biological Research-Thessaloniki*, *27*(1), 1–10.

Capra, F., & Luisi, P. L. (2014). *The systems view of life: A unifying vision*. Cambridge University Press.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist*, *84*(2), 242–264.

Cartwright, N., & Jones, M. (1991). How to hunt quantum causes. *Erkenntnis*, *35*(1), 205–231.

Chalmers, D. J. (1995). On implementing a computation. *Minds and Machines*, *4*, 391–402.

Chapman, S. (1968). Catching a baseball. *American Journal of Physics*, *36*, 868–870.

Chemero, A. (2011). *Radical embodied cognitive science* (First MIT Press paperback edition). MIT Press.

Chemero, A., & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, *75*(1), 1–27.

Chirimuuta, M. (2020). The reflex machine and the cybernetic brain: The critique of abstraction and its application to computationalism. *Perspectives on Science*, *28*(3), 421–457.

Christensen, W. (2012). Natural Sources of Normativity. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43, 104–112.

Christensen, W. D., & Hooker, C. A. (2000). Autonomy and the emergence of intelligence: Organised interactive construction. *Communication and Cognition—Artificial Intelligence*, *17*(3–4), 133–157.

Christensen, W. D., & Bickhard, M. H. (2002). The process dynamics of normative function. *The Monist*, *85*(1), 3–28.

Ciaunica, A., Constant, A., Preissl, H., & Fotopoulou, K. (2021). The first prior: From co-embodiment to co-homeostasis in early life. *Consciousness and cognition*, *91*, 103117.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, *53*, 3–27.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 3*. MIND Group. https://doi.org/10.15502/9783958573031

Clark, A. (2019). Consciousness as generative entanglement. *The Journal of Philosophy*, *116*(12), 645–662.

Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese*, *101*(3), 401–431.

Clegg, J. S. (2001). Cryptobiosis—A peculiar state of biological organization. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *128*(4), 613–624.

Collier, J. (2004). Self-organisation, individuation and identity. *Revue Internationale de Philosophie*, *59*, 151–172.

Collier, J. (2008). A dynamical account of emergence. *Cybernetics & Human Knowing*, *15*(3–4), 75–86.

Colombetti, G. (2010). Enaction, Sense-Making, and Emotion. In J. Stewart, O. Gapenne, E. A. Di Paolo (Eds.), *Enaction: Toward a New Paradigm for Cognitive Science*. (pp. 145–164). MIT Press.

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind*. MIT press.

Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*, *198*(14), 3463–3488.

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*(5), 1–26.

Constant, A. (2021). The free energy principle: It's not about what it takes, it's about what took you there. *Biology & Philosophy*, *36*(2), 10.

Constant, A., Ramstead, M. J., Veissière, S. P., & Friston, K. J. (2019). Regimes of expectations: An active inference model of social conformity and human decision making. *Frontiers in Psychology*, *10*, 679.

Corcoran, A.W., & Hohwy, J. (2018). Allostasis, interoception, and the free energy principle: Feeling our way forward. In M. Tsakiris & H. De Preester (Eds.), *The* *interoceptive mind: From homeostasis to awareness* (Chap. 15, pp. 272–292). Oxford University Press.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 1–45.

Costanzo, J. P., do Amaral, M. C. F., Rosendale, A. J., & Lee Jr, R. E. (2013). Hibernation physiology, freezing adaptation and extreme freeze tolerance in a northern population of the wood frog. *Journal of Experimental Biology*, *216*(18), 3461–3473.

Crane, T. (2015). *The mechanical mind: A philosophical introduction to minds, machines and mental representation*. Routledge.

Da Costa, L., Friston, K., Heins, C., & Pavliotis, G. A. (2021). Bayesian mechanics for stationary processes. *Proceedings of the Royal Society A*. 477:20210518. 1–26

Da Costa, L., Parr, T., Sengupta, B., & Friston, K. J. (2021). Neural dynamics under active inference: Plausibility and efficiency of information processing. *Entropy*, *23*(4), 454.

Henderson, David. (2002). Norms, normative principles, and explanation. *Philosophy of the Social Sciences*, *32*, 329–364.

Davidson, D. (1963). Actions, reasons and causes. *The Journal of Philosophy*, *60*(23), 685–700.

Dawkins, R. (1986). *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. W. W. Norton & Company.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation*, *7*(5), 889–904.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507.

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, *4*(4), 429–452. https://doi.org/10.1007/s11097-005-9002-y

Di Paolo, E. A. (2010). Overcoming autopoiesis: An enactive detour on the way from life to society. In R. Magalhães, & R. Sanchez (Eds.), *Advanced series in management*. (pp. 43–68) Emerald Group Publishing Limited.

Di Paolo, E. (2009). Extended life. *Topoi*, *28*, 9–21.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.

Di Paolo, E. A., Cuffari, E. C., & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. MIT press.

Di Paolo, E., & Thompson, E. (2014). The enactive approach. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 86–96). Routledge.

Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences, 3*. 1–39.

Di Paolo, E., Buhrmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.

DiFrisco, J. (2014). Hylomorphism and the metabolic closure conception of life. *Acta Biotheoretica*, *62*(4), 499–525.

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese*, *195*(2), 5115–5139.

Dressler, W. W. (1999). Modernization, stress, and blood pressure: New directions in research. *Human Biology*, *71*(4), 583–605.

Dretske, F. (1993). The nature of thought. *Philosophical Studies*, *70*(2), 185–199.

Dreyfus, H. L. (2002). Intelligence without representation–Merleau-Ponty's critique of mental representation The relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1*(4), 367–383.

Dupré, J., & Nicholson, D. (2018). A manifesto for a processual philosophy of biology. In D. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (pp. 3–45). Oxford University Press.

Dupuy, J. P. (2018). Cybernetics is an antihumanism. Technoscience and the rebellion against the human condition. In S. Loeve, X. Guchet, & B. Bensaude Vincent, (Eds.) *French philosophy of technology* (pp. 139–156). Springer.

Earman, J., & Rédei, M. (1996). Why ergodic theory does not explain the success of equilibrium statistical mechanics. *British Journal for the Philosophy of Science*, *47*(1), 63–78.

Egan, D., Reynolds, S., & Wendland, A. J. (Eds.). (2013). *Wittgenstein and Heidegger*. Routledge.

Enoch, David. (2011). *Taking morality seriously: A defense of robust realism*. Oxford University Press.

Evans, G. (1982). *The varieties of reference*. Clarendon Press.

Evans, D. J., & Searles, D. J. (2002). The fluctuation theorem. *Advances in Physics*, *51*(7), 1529–1585.

Eyer, J. (1975). Hypertension as a disease of modern society. *International Journal of Health Services, 5*(4), 539–558.

Field, H. (2003). Causation in a physical world. *Oxford handbook of metaphysics* (pp. 435–460).

Fiorillo, C. D. (2010). A neurocentric approach to Bayesian inference. *Nature Reviews Neuroscience*, *11*(8), 605–605.

Firth, N., Jensen, S. O., Kwong, S. M., Skurray, R. A., & Ramsay, J. P. (2018). Staphylococcal plasmids, transposable and integrative elements. *Microbiology Spectrum*, *6*(6), 6.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, *8*, 457.

Fleischaker, G. R. (1988). Autopoiesis: The status of its system logic. *BioSystems*, *22*(1), 37–49.

Fleischaker, G. R. (1990). Origins of life: An operational definition. *Origins of Life and Evolution of the Biosphere, 20*(2), 127–137.

Foglia, L., & Grush, R. (2011). of a Purely Enactive (Non-Representational) Account of Imagery. *Journal of Consciousness Studies*, *18*(5–6), 35–43.

Friston, K. J. (2000). The labile brain. II. Transients, complexity and selection. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *355*(1394), 237–252.

Friston, K. J. (2002). Functional integration and inference in the brain. *Progress in Neurobiology*, *68*(2), 113–143.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of The Royal Society Interface*, *10*(86), 20130475.

Friston, K. J. (2018). Am I self-conscious? (Or does self-organization entail self-consciousness?). *Frontiers in Psychology*, *9*, 579.

Friston, K. J. (2019a). A free energy principle for a particular physics. *arXiv preprint arXiv:1906.10184*.

Friston, K. J. (2019b). Beyond the desert landscape. *Andy Clark and His Critics,* (pp. 174–190). Oxford University Press

Friston, K. J., & Ao, P. (2012). Free energy, value, and attractors. *Computational and Mathematical Methods in Medicine*, *2012*. 1–27.

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2022). The free energy principle made simpler but not too simple (arXiv:2201.06387). arXiv. http://arxiv.org/abs/2201.06387.

Friston, K., Da Costa, L., Sakthivadivel, D. A. R., Heins, C., Pavliotis, G. A., Ramstead, M., & Parr, T. (2023). Path integrals, particular kinds, and strange things. *Physics of Life Reviews*, *47*, 35–62.

Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2021). Parcels and particles: Markov blankets in the brain. *Network Neuroscience*, *5*(1), 211–251.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews*, *68*, 862–879.

Friston, K. J., FitzGerald, T., Rigoli, F., Schawartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*, 1–49.

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1–3), 70–87.

Friston, K. J., & Mathys, C. (2016). I am therefore I think. In M. Leuzinger-Bohleber, Arnold, S., & Solms, M. (Eds.). (2016). *The Unconscious: A bridge between psychoanalysis and cognitive neuroscience* (p. 113). Taylor and Francis.

Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, *34*(1), 220–234.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.

Friston, K. J., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*, 130.

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, *22*(5), 516.

Froese, T., & Stewart, J. (2010). Life after Ashby: Ultrastability and the autopoietic foundations of biological autonomy. *Cybernetics & Human Knowing*, *17*(4), 7–49.

Fulda, F. C. (2017). Natural agency: The case of bacterial cognition. *Journal of the American Philosophical Association*, *3*(1), 69–90.

Gallagher, S. (2017). *Enactivist interventions: Rethinking the mind.* Oxford University Press.

Gallagher, S. (2018). Rethinking Nature: Phenomenology and a Non-reductionist Cognitive Science. *Australasian Philosophical Review*, *2*(2), 125–137.

Gallavotti G. (1999). *Statistical mechanics: A short treatise*. Springer.

Gangopadhyay, N., & Kiverstein, J. (2009). Enactivism and the unity of perception and action. *Topoi*, *28*(1), 63–73.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *Neurons, Behavior, Data Analysis, and Theory*.

Gershman, S. J., & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. I. Rabinovich, K. J. Friston, & P. Varona (Eds.), *Principles of brain dynamics: Global state interactions* (pp. 293–312). MIT Press.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton-Mifflin.

Gier, N. F. (1981). *Wittgenstein and phenomenology: A comparative study of the later Wittgenstein, Husserl, Heidegger, and Merleau-Ponty*. State University of New York Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559–582.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(1), S342–S353.

Glymour, C. (1999). Rabbit hunting. *Synthese, 121*(1/2), 55-78.

Glymour, C. (2006). Markov properties and quantum experiments. In W. Demopoulos & I. Pitowsky (Eds.), *Physical theory and its interpretation* (pp. 117–126). Springer.

Glymour, C. (2010). What is right with "Bayes net methods" and what is wrong with "hunting causes and using them"? *The British Journal for the Philosophy of Science*, *61*(1), 161–211.

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *The Journal of Philosophy*, *113*(10), 481–506.

Good, I. J. (1959). A theory of causality. *British Journal for the Philosophy of Science*, *9*, 307–310.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B*, *290*(1038), 181–197. [aAC, KF].

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396.

Grush, R. (2007). Skill theory v2. 0: Dispositions, emulation, and spatial perception. *Synthese*, *159*(3), 389–416.

Gu, X., & FitzGerald, T. H. B. (2014). Interoceptive inference: Homeostasis and decision-making. *Trends in Cognitive Sciences*, *18*(6), 269–270.

Gualeni, S. (v 1.1, 2017). *Something something soup something*. Digital game developed by I. Kniestedt, R. Fassone, and J. Schellekens. http://soup.gua-le-ni.com.

Guay, Alexandre & Pradeu, Thomas (2016). To Be Continued: The Genidentity of Physical and Biological Processes. In A. Guay & T. Pradeu (Eds.), *Individuals Across the Sciences*. Oxford University Press. pp. 317-347.

Haldane, J. S. (1917). *Organism and environment, as illustrated by the physiology of breathing*. Yale University Press.

Hammer, T. J., Sanders, J. G., & Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiology Letters*, *366*(10), 117.

Harvey, K. (2018). An open-ended approach to Piagetian development of adaptive behavior. *Open Access Library Journal, 5*(03), 1.

Hatfield, G. (1984). Spatial perception and geometry in Kant and Helmholtz. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1984*(2), 569–580.

Hatfield, G. C. (1990). *The natural and the normative: Theories of spatial perception from Kant to Helmholtz*. MIT Press.

Hatfield, G. C. (1990). *The natural and the normative: Theories of spatial perception from Kant to Helmholtz*. MIT Press.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. Stich, & J. Garron (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Lawrence Erlbaum.

Hausman, D. M. (2005). Causal relata: Tokens, types, or variables? *Erkenntnis*, *63*(1), 33–54.

Hediger, M. A., Johnson, D. F., Nierlich, D. P., & Zabin, I. (1985). DNA sequence of the lactose operon: The lacA gene and the transcriptional termination region. *Proceedings of the National Academy of Sciences*, *82*(19), 6414–6418.

Helmholtz, H. (1962). Concerning the perceptions in general. In *Treatise on physiological optics* (3rd ed., Vol. III, J. Southall, trans.). Dover. Original work published 1866.

Henderson, D. (2010). Explanation and rationality naturalized. *Philosophy of the Social Sciences*, *40*(1), 30–58.

Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*(11), 1315–1326.

Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. J. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Y. Georgiev, and M. Ramstead, J. M. Smart, C. L. F. Martinez, and M. E. Price (Eds.), *Evolution, development and complexity* (pp. 195–227). Springer.

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, *33*(2), 398–446.

Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of Science*, *43*(2), 290–292.

Hinton, G. E., & Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on computational learning theory*, *6*, 5–13.

Hipólito, I., Ramstead, M. J., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2021). Markov blankets in the brain. *Neuroscience & Biobehavioral Reviews*, *125*, 88–97.

Hobson, A. (2013). There are no particles, there are only fields. *American Journal of Physics*, *81*(3), 211–223.

Hoffmann, P. M. (2012). *Life's ratchet: How molecular machines extract order from chaos*. Basic Books.

Howarth, N., Odnoletkova, N., Alshehri, T., Almadani, A., Lanza, A., & Patzek, T. (2020). Staying cool in a warming climate: Temperature, electricity and air conditioning in Saudi Arabia. *Climate*, *8*(1), Article 1.

Hohwy, J. (2013). *The predictive mind*. OUP Oxford.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259–285.

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing* (p. 2). MIND Group. https://doi.org/10.15502/9783958573048.

Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese*, *199*(1), 29–53.

Humphreys, S., Deyermond, R., Bali, I., Stevenson, M., & Fee, J. P. H. (2005). The effect of high altitude commercial air travel on oxygen saturation. *Anaesthesia*, *60*(5), 458–460.

Hurley, S. (1998). *Consciousness in action*. Harvard University Press.

Hurley, S. (2003). Animal action in the space of reasons. *Mind & Language*, *18*(3), 231–257.

Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. *Behavioral and Brain Sciences*, *31*(1), 1–22.

Husbands, P., & Holland, O. (2012). Warren McCulloch and the British cyberneticians. *Interdisciplinary Science Reviews*, *37*(3), 237–253.

Husserl, E. (1970). *The crisis of European sciences and transcendental phenomenology. An introduction to phenomenology* (D. Carr, Trans.). Northwestern University Press. Original work published 1936.

Husserl, E. (1982). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy. First book: General introduction to a pure phenomenology* (F. Kerstens, Trans.). Kluwer. Original work published 1913.

Husserl, E. (1997). *Thing and space: Lectures of 1907* (R. Rojcewicz, Trans.). Kluwer. Original work published in 1907.

Husserl, E. (1999). *Cartesian meditations: An introduction to phenomenology* (D. Cairns, Trans.). Martinus Nijhoff. Original work published 1929.

Husserl, E. (2001). *Analyses concerning passive and active synthesis*. Lectures on Transcendental Logic. (A. J. Steinbock, Trans.). Kluwer Academic Publishers. Original work published 1920.

Hutto, D. D., & Myin, E. (2012). *Radicalizing enactivism: Basic minds without content.* MIT Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT Press.

Jacob, P. (2019). Intentionality, *The Stanford encyclopedia of philosophy*, Edward N. Zalta (Ed.), https://plato.stanford.edu/archives/win2019/entries/intentionality/

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, *3*(3), 318–356.

James, W. (1890). *The principles of psychology, Vols. I, II*. Harvard University Press.

Jebari, J. (2019). Empirical moral rationalism and the social constitution of normativity. *Philosophical Studies*, *176*(9), 2429–2453.

Jeffrey, R. C. (1969). Statistical explanation vs. statistical inference. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel* (pp. 104–113). Springer.

Jiang, R., Li, W., Lu, X. X., Xie, J., Zhao, Y., & Li, F. (2021). Assessment of temperature extremes and climate change impacts in Singapore, 1982–2018. *Singapore Journal of Tropical Geography*, *42*(3), 378–396.

Jonas, H. (1953). A critique of cybernetics. *Social Research*, *20(2),* 172–192.

Jonas, H. (1966). *The phenomenon of life: Toward a philosophical biology*. Northwestern University Press.

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

Kanai, R., Komura, Y., Shipp, S., & Friston, K. J. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1668), 20140169.

Kant, I. (2008). *Groundwork for the Metaphysics of Morals*. Yale University Press. Original work published 1785.

Kaplan, D. M. (2017). Mechanisms and dynamical systems. In S. Glennan & Illari, P. (Eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 267–280). Routledge.

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science*, *3*(2), 438–444.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.

Kauffman, S. A. (1986). Autocatalytic sets of proteins. *Journal of Theoretical Biology*, *119*(1), 1–24.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press.

Kauffman, S. A. (2000). *Investigations*. Oxford University Press.

Kauffman, S. A. (2019). *A world beyond physics: The emergence and evolution of life*. Oxford University Press.

Keilin, D. (1959). The Leeuwenhoek Lecture-The problem of anabiosis or latent life: History and current concept. *Proceedings of the Royal Society of London. Series B-Biological Sciences*, *150*(939), 149–191.

Keller, E. F. (2008). Organisms, machines, and thunderstorms: A history of self-organization, part one. *Historical Studies in the Natural Sciences*, *38*(1), 45–75.

Keller, E. F. (2009). Organisms, machines, and thunderstorms: A history of self-organization, part two. Complexity, emergence, and stable attractors. *Historical Studies in the Natural Sciences*, *39*(1), 1–31.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*(2), 424–435.

Kenny, Anthony. (1989). *The metaphysics of mind*. Clarendon Press.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387–2415.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387–2415. https://doi.org/10.1007/s11229-017-1435-7.

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. In S. Robins, J. Symons, & P. Calvo (Eds.), *The Routledge companion to philosophy of psychology* (pp. 384–409). Routledge.

Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese, 198*(5), 4791–4810.

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, *21*(2), 264–281.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, *15*(138), 20170792. http://doi.org/10.1098/rsif.2017.0792.

Korbak, T. (2021). Computational enactivism under the free energy principle. *Synthese*, *198*, 2743–2763. https://doi.org/10.1007/s11229-019-02243-4.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.

Korsgaard, C. M. (2018a). *Fellow creatures: Our obligations to the other animals*. Oxford University Press.

Korsgaard, C. M. (2018b). Rationality. In L. Gruen (Ed.), *Critical Terms for Animal Studies*, (pp. 294–307). University of Chicago Press.

Koutroufinis, S. A. (2017). Organism, machine, process. Towards a process ontology for organismic dynamics. *Organisms. Journal of Biological Sciences*, *1*(1), 23–44.

Kuchling, F., Friston, K., Georgiev, G., & Levin, M. (2020). Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems. *Physics of life reviews*, *33*, 88–108.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. 1–101

Langton, C. G. (1989). Artificial life. In C. G. Langton (Ed.), *Artificial life* (pp.1–47). Routledge.

Lawes, C. M., Vander Hoorn, S., Law, M. R., Elliott, P. A. U. L., MacMahon, S., & Rodgers, A. (2004). *High blood pressure. Comparative quantification of health risks: Global and regional burden of disease attributable to selected major risk factors* (pp. 281–390). World Health Organization.

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of Optical Society of America A*, *20*(7), 1434–1448. [aAC].

Letelier, J. C., Cárdenas, M. L., & Cornish-Bowden, A. (2011). From L'Homme machine to metabolic closure: Steps towards understanding life. *Journal of Theoretical Biology*, *286*, 100–113.

Lewin, K. (1922). *Der Begriff der Genese in Physik, Biologie und Entwicklungsgeschichte: Eine Untersuchung zur vergleichenden Wissenschaftslehre*. Springer.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, *103*(412), 473–490.

Longo, G., & Montévil, M. (2013). Extended criticality, phase spaces and enablement in biology. *Chaos, Solitons & Fractals*, *55*, 64–79.

Longo, G., & Montévil, M. (2014). From physics to biology by extending criticality and symmetry breakings. In *Perspectives on Organisms*, (pp. 161–185). Springer.

Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, 1379–1392.

Lotze, R. H. (1852). *Medizinische Psychologie oder Physiologie der Seele* (pp. 287–325). WeidmannÕsche Buch-handlung.

Lusthaus, D. (2014). *Buddhist Phenomenology: A philosophical investigation of Yogacara Buddhism and the Ch'eng Wei-shih Lun*. Routledge.

Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing*, *7*(1), 11–29.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25.

Madary, M. (2012). How would the world look if it looked as if it were encoded as an intertwined set of probability density distributions? *Frontiers in Psychology*, *3*, 419.

Maddy, P. (1990). Physicalistic platonism. In A.D. Irvine (Ed.), *Physicalism in mathematics* (pp. 259–289). Springer.

Maddy, P. (1997). *Naturalism in mathematics*. Clarendon Press.

Mann, S. F., Pain, R., & Kirchhoff, M. D. (2022). Free energy: A user's guide. *Biology & Philosophy*, *37*(4), 33.

Matthen, M. (2014). Debunking enactivism: A critical notice of Hutto and Myin's radicalizing enactivism. *Canadian Journal of Philosophy*, *44*(1), 118–128.

Matthen, M., & Ariew, A. (2002). Two ways of thinking about fitness and natural selection. *The Journal of Philosophy*, *99*(2), 55–83.

Maturana, H. R. (1970). *Biology of cognition* (pp. 4–9). Biological Computer Laboratory, Department of Electrical Engineering, University of Illinois.

Maturana, H. R. (1975). The organization of the living: A theory of the living organization. *International Journal of Man-Machine Studies*, *7*(3), 313–332.

Maturana, H. (1987). Everything is said by an observer. In W. I. Thompson (Ed.), *GAIA: A way of knowing* (pp. 65–82). Hudson, N.Y.: Lindisfarne Press.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living* (pp. 59–140). Kluwer Academic. Original work published 1972.

Maynard, C., & Weinkove, D. (2020). Bacteria increase host micronutrient availability: Mechanisms revealed by studies in C. elegans. *Genes & Nutrition*, *15*(1), 1–11.

McDowell, J. (1994). *Mind and world*. Harvard University Press.

McEwen, B. S. (2000). The neurobiology of stress: From serendipity to clinical relevance. *Brain Research*, *886*(1–2), 172–189.

Mehler, D. M. A., & Kording, K. P. (2018). The lure of causal statements: Rampant mis-inference of causality in estimated connectivity. *arXiv e-prints*, arXiv-1812.

Meincke, A. S. (2019). Autopoiesis, biological autonomy and the process view of life. *European Journal for Philosophy of Science*, *9*(1), 1–16.

Menary, R., & Gillett, A. J. (2021). Are Markov blankets real and does it matter? In D. Mendonça, M. Curado, & S. S. Gouveia (Eds.), *The philosophy and science of predictive processing* (pp. 39–58). Bloomsbury Academic.

Merleau-Ponty, M. (1963). *The structure of behavior* (A. Fisher, Trans.). Dusquene Lniversity Press. Original work published 1942.

Merleau-Ponty, M. (1964). *The primacy of perception: And other essays on phenomenological psychology, the philosophy of art, history, and politics*. Northwestern University Press.

Merleau-Ponty, M. (2004). Eye and Mind. In T. Baldwin (Ed.), *Maurice Merleau-Ponty: Basic writing* (pp. 291–234). Routledge. Original work published 1961.

Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. Landes, Trans.). Routledge. Original work published 1945.

Meyer, R., & Brancazio, N. (2022). Putting down the revolt: Enactivism as a philosophy of nature. *Frontiers in Psychology*, *13*, 948733.

Mi, S., Lee, X., Li, X. P., Veldman, G. M., Finnerty, H., Racie, L., & McCoy, J. M. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, *403*(6771), 785–789.

Millidge, B., Seth, A., & Buckley, C. L. (2021a). A mathematical walkthrough and discussion of the free energy principle. *arXiv preprint arXiv:2108.13343*.

Millidge, B., Tschantz, A., & Buckley, C. L. (2021b). Whence the expected free energy? *Neural Computation*, *33*(2), 447–482. https://doi.org/10.1162/neco_a_01354.

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

Mohanty, J. N. (1978). Husserl's transcendental phenomenology and essentialism. *The Review of Metaphysics*, *32*(3), 299–321.

Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, *372*, 179–191.

Moran, D. (2002). *Introduction to phenomenology*. Routledge.

Moreno, A., & Ruiz-Mirazo, K. (1999). Metabolism and the problem of its universalization. *BioSystems*, *49*(1), 45–61.

Moreno, A., & Mossio, M. (2015). *Biological autonomy. A philosophical and theoretical enquiry*. Springer.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators* (p. 347). Cambridge University Press.

Morgan, W. (2021). Are organisms substances or processes? *Australasian Journal of Philosophy*, *100(3)*, 1–15.

Mossio, M., & Bich, L. (2017). What makes biological organisation teleological? *Synthese*, *194*(4), 1089–1114.

Mossio, M., Saborido, C., and Moreno, A. (2009). An Organizational Account of Biological Functions. *British Journal for the Philosophy of Science*, 60: 813–41.

Murata, R., Nozu, R., Mushirobira, Y., Amagai, T., Fushimi, J., Kobayashi, Y., & Nakamura, M. (2021). Testicular inducing steroidogenic cells trigger sex change in groupers. *Scientific Reports*, *11*(1), 1–7.

Musmann, H. (1979). Predictive image coding. In W. K. Pratt (Ed.), *Image transmission techniques* (pp. 73–112), Academic Press.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Nagel, T. (2012). *Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false*. Oxford University Press.

Nanay, Bence (2014). Empirical problems with anti-representationalism. In B. Brogaard (ed.), *Does Perception have Content?* (pp. 39–50), Oxford University Press.

Nave, K., Deane, G., Miller, M., & Clark, A. (2020). Wilding the predictive brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(6), e1542.

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.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Springer, Dordrecht.

Netburn, D. (2014). In Alaska, wood frogs freeze for seven months, thaw and hop away. *Los Angeles Times*. Retrieved from https://www.latimes.com/science/sciencenow/la-sci-sn-alaskan-frozen-frogs-20140723-story.html.

Neuman, Y. (2006). Cryptobiosis: A new theoretical perspective. *Progress in Biophysics and Molecular Biology*, *92*(2), 258–267.

Nicholson, D. (2018). Reconceptualizing the organism: From complex machine to flowing stream. In D. Nicholson & J. Dupré (Eds.), *Everything flows: Towards a processual philosophy of biology* (pp. 139–166). Oxford University Press.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(1), 152–163.

O'regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, *24*(5), 939–973.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process.* Oxford University Press.

Orlandi, N. (2018). Predictive perceptual systems. *Synthese*, *195*(6), 2367–2386.

Overgaard, S. (2006). The problem of other minds: Wittgenstein's phenomenological perspective. *Phenomenology and the Cognitive Sciences*, *5*(1), 53–73.

Palacios P. (2018). Had we but world enough, and time . . . but we don't!: Justifying the thermodynamic and infinite-time limits in statistical mechanics. *Foundations of Physics*, *48*(5), 526–541.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. J. (2017). Biological self-organisation and Markov blankets. *BioRxiv*. Preprint. https://www.biorxiv.org/content/10.1101/227181v1.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., & Friston, K. J. (2020). On Markov blankets and hierarchical self-organisation. *Journal of Theoretical Biology*, *486*, 110089.

Palmer, C. J., Seth, A. K., & Hohwy, J. (2015). The felt presence of other minds: Predictive processing, counterfactual predictions, and mentalising in autism. *Consciousness and Cognition*, *36*, 376–389.

Papineau, D. (1992, January). Can we reduce causal direction to probabilities? In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1992, No. 2, pp. 238–252). Philosophy of Science Association.

Papineau, D. (2022). The statistical nature of causation. *The Monist*, *105*(2), 247–275.

Parfit, Derek. (2006). Normativity. In R. Shafer-Landau (Ed.), *Oxford studies in metaethics* (pp. 325–380). Clarendon Press.

Parfit, Derek (2011). *On What Matters, vol. 2*. Oxford University Press.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*(1), 1–21.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: The free energy principle in mind, brain, and behavior*. The MIT Press.

Pattee, H. H. (1982). *Cell psychology: An evolutionary approach to the symbol-matter problem*. Reprinted in H. H. Pattee & J. Rączaszek-Leonardi. *Laws, Language and Life*, (pp. 165–179). Dordrecht: Springer, 2012.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2001). Bayesianism and causality, or, why I am only a half-Bayesian. In D. Corfield & J. Williamson (Eds.) *Foundations of Bayesianism* (pp. 19–36). Kluwer Academic Publishers.

Pedersen, B. H., Malte, H., Ramløv, H., & Finster, K. (2020). A method for studying the metabolic activity of individual tardigrades by measuring oxygen uptake using microrespirometry. *Journal of Experimental Biology*, *223*(22), jeb233072.

Peterson, E. L. (2017). *The life organic: The theoretical biology club and the roots of epigenetics*. University of Pittsburgh Press.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger, & Wiese W (Eds.), *Philosophy and predictive processing*. MIND Group.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences*, *22*(4), 294–306.

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. Edinburgh University Press, 1967a (Trans. B. Walsh). Original work published in 1967.

Piccinini, G. (2020). Mechanisms, multiple realizability, and medium independence. *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Pickering, A. (2010). *The cybernetic brain*. University of Chicago Press.

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.

Powers, W. T. (1973). *Behavior: The control of perception*. Aldine.

Price, H. (1992, January). The direction of causation: Ramsey's ultimate contingency. In *PSA: Proceedings of the biennial meeting of the philosophy of science association* (Vol. 1992*,* No. 2, pp. 253–267). Philosophy of Science Association.

Price, J. C. (1979). Assessment of the urban heat island effect through the use of satellite data. *Monthly Weather Review*, *107*(11), 1554–1557.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos*. New Science Library.

Prigogine, I., & Stengers, I. (1997). *The end of certainty*. Simon and Schuster.

Proksch, S (2021, October 15–17). *Acoustic social worlds; from Markov blankets to interpersonal synergies* [Conference presentation]. Cognitio 2021, L'Université du Québec à Montréal, Canada. https://www.youtube.com/watch?v=CiiPVa3yu6s.

Putnam, H. (1975). The mental life of some machines. *Mind, language and reality, philosophical papers, volume 2* (pp. 408–428). Cambridge University Press.

Raja, V., Valluri, D., Baggs, E., Chemero, A., & 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.

Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge University Press.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, *24*, 1–16.

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

Ramstead, M. J., Hesp, C., Tschantz, A., Smith, R., Constant, A., & Friston, K. J. (2021). Neural and phenotypic representation under the free-energy principle. *Neuroscience & Biobehavioral Reviews*, *120*, 109–122.

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239.

Ramstead, M. J. D., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: Beyond internalism and externalism. *Synthese*, *24*(6), 17.

Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*, *13*(3), 20220029.

Raposo de Magalhães, C., Schrama, D., Nakharuthai, C., Boonanuntanasarn, S., Revets, D., Planchon, S., & Rodrigues, P. M. (2021). Metabolic plasticity of gilthead seabream under different stressors: Analysis of the stress responsive hepatic proteome and gene expression. *Frontiers in Marine Science*, *8*, 469.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Reichenbach, H. (1956). *The direction of time* (Vol. 65). University of California Press.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.), *The Oxford handbook of philosophy of perception* (pp. 694–716). Oxford University Press.

Roelofs, L. (2018). Why imagining requires content: A reply to a reply to an objection to radical enactive cognition. *Thought: A Journal of Philosophy*, *7*(4), 246–254.

Rosas, F. E., Mediano, P. A., Biehl, M., Chandaria, S., & Polani, D. (2020,). Causal blankets: Theory and algorithmic framework. In Verbelen, T., Lanillos, P., Buckley, C. L., & De Boom, C. (Eds.), *Active Inference*. IWAI 2020. Communications in Computer and Information Science (Vol. 1326, pp. 187–198). Springer.

Rosen, R. (1991). *Life itself: A comprehensive inquiry into the nature, origin, and fabrication of life*. Columbia University Press.

Rosen, R. (1999). *Essays on life itself*. Columbia University Press.

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science, 82*(1), 32–54.

Rousseau, J. J. (2018). *Rousseau: The social contract and other later political writings* (V. Gourevitch, Trans.). Cambridge University Press. Original work published in 1762.

Roy, J. M., Petitot, J., Pachoud, B., & Varela, F. J. (1999). Beyond the gap: An introduction to naturalizing phenomenology. In J. Petitot, F. J. Varela, B. Pachoud, J. M. Roy (Eds.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 1–83). Stanford University Press.

Rubin, S., Parr, T., Da Costa, L., & Friston, K. J. (2020). Future climates: Markov blankets and active inference in the biosphere. *Journal of the Royal Society Interface*, *17*(172), 20200503.

Ruiz-Mirazo, K., & Mavelli, F. (2007). Simulation model for functionalized vesicles: Lipid-peptide integration in minimal protocells. In F. A. e Costa, L. M. Rocha, E. Costa, I. Harvey, & A. Coutinho (Eds.), *ECAL* (Vol. 4648, pp. 32- 41). Springer.

Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial life*, *10*(3), 235–259.

Ruiz-Mirazo, K., Peretó, J., & Moreno, A. (2004). A universal definition of life: Autonomy and open-ended evolution. *Origins of Life and Evolution of the Biosphere*, *34*(3), 323–346.

Rutar, D., de Wolff, E., van Rooij, I., & Kwisthout, J. (2022). Structure learning in predictive processing needs revision. *Computational Brain & Behavior*, *5*(2), 234–243.

Sakthivadivel, D. A. R. (2022). Regarding flows under the free energy principle: A comment on "how particular is the physics of the free energy principle?" by Aguilera, Millidge, Tschantz, and Buckley. *Physics of Life Reviews*, *42*, 25–28.

Salmon, W. (1998). Probabilistic causality. In *Causality and explanation* (pp. 208–232). Oxford University Press. (Original work published 1980 in *Pacific Philosophical Quarterly*, 61, 50–74).

Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford University Press, USA.

Schaffer, J. (2016). The Metaphysics of Causation, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/.

Schellenberg, S. (2007). Action and self-location in perception. *Mind*, *116*(463), 603–632.

Schrödinger, E. (1951). *What is life? The physical aspect of the living cell*. At the University Press.

Schütz, A. (1959). Type and eidos in Husserl's late philosophy. *Philosophy and Phenomenological Research*, *20*(2), 147–165.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. J. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, *4*, 710.

Sellars, W. (1956). Empiricism and the philosophy of mind. *Minnesota Studies in the Philosophy of Science*, *1*(19), 253–329.

Sender, R., & Milo, R. (2021). The distribution of cellular turnover in the human body. *Nature Medicine*, *27*(1), 45–48.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5*(2), 97–118.

Seth, A. K. (2015). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience*, *6*(2–3), 111–117.

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious inferences: Reply to sun and firestone on the dark room problem. *Trends in Cognitive Sciences*, *24*(9), 681–683.

Shakhashiri, B. Z. (1983). *Chemical demonstrations: A handbook for teachers of chemistry, Volume 1*. The University of Wisconsin Press.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shi, Yun Q., & Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. CRC Press.

Sims, M. (2021). How to count biological minds: Symbiosis, the free energy principle, and reciprocal multiscale integration. *Synthese*, *199*(1), 2157–2179.

Smith, F. A., Elliott Smith, R. E., Lyons, S. K., & Payne, J. L. (2018). Body size downgrading of mammals over the late Quaternary. *Science*, *360*(6386), 310–313.

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

Spelke, E. S., & Kinzler, K. D. (2009). Innateness, learning, and rationality. *Child Development Perspectives*, *3*(2), 96–98.

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In Maria Carla Galavotti (Ed.), *Stochastic causality* (pp. 157–172). CSLI Publ. ISBN 1-57586-321-9.

Sprevak, M. (2018). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 175–191). Routledge.

Sprevak, Mark. (2021). *Predictive coding I: Introduction*. [Preprint].

Stafford, J. (2005). *A philosophical interpretation of Judea Pearl's theory of causality* (Doctoral dissertation, University of Tasmania).

Steinhauser, M. L., Olenchock, B. A., O'Keefe, J., Lun, M., Pierce, K. A., Lee, H., & Fazeli, P. K. (2018). The circulating metabolome of human starvation. *JCI Insight*, *3*(16), 1–16.

Stenmark, M. (2001). Evolution, purpose and God. *Ars Disputandi*, *1*(1), 44–52.

Sterling P. (2004). Principles of allostasis: optimal design, predictive regulation, pathophysiology and rational therapeutics. In J. Schulkin (Ed.), *Allostasis* (pp. 17–64). MIT Press.

Sterling, P., & Eyer, J. (1988). Allostasis: A new paradigm to explain arousal pathology. In S. Fisher & J. Reason (Eds.), *Handbook of life stress, cognition and health* (pp. 629–649). John Wiley & Sons.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., Tittgemeyer, M., Fleming, S.M., Haker, H., Seth, A.K. & Petzschner, F. H. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in human neuroscience*, *10*, 550.

Stepp, N., Chemero, A., & Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science*, *3*(2), 425–437.

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*, *24*(5), 346–348.

Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland Publishing Company.

Swenson, R. (1989). Emergent attractors and the law of maximum entropy production: Foundations to a theory of general evolution. *Systems research*, *6*(3), 187–197.

Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Thompson, E. (2011). Reply to commentaries. *Journal of Consciousness Studies*, *18*(5–6), 176–223.

Thompson, E. (2018). Review of evolving enactivism: Basic minds meet content. *Notre Dame Philosophical Reviews*. https://ndpr.nd.edu/news/evolving-enactivism-basic-minds-meet-content/.

Toyama, B. H., & Hetzer, M. W. (2013). Protein homeostasis: Live long, won't prosper. *Nature Reviews Molecular Cell Biology*, *14*(1), 55–61.

Turing, A. M. (1948). "Intelligent machinery," National physical laboratory report. In B. Meltzer, & D. Michie (Eds.), *Machine intelligence* (p. 5). Edinburgh University Press (1969).

van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, *29*(3), 315–329.

van Es, T., & Hipólito, I. (2020). Free-energy principle, computationalism and realism: A tragedy. Philsci-archive. Preprint.

van Fraassen, B. (1980). *The scientific image*. The Clarendon Press.

van Fraassen, B. (1982). The Charybdis of realism: Epistemological implications of Bell's inequality. *Synthese*, *52*, 25–38.

Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.

van Lith, J. (2001). Ergodic theory, interpretations of probability and the foundations of statistical mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *32*(4), 581–594.

Varela, F. J, (1979). *Principles of biological autonomy*. North-Holland.

Varela, F. J. (2011). Preface to the second edition of "De Máquinas y Seres Vivos—Autopoiesis: La organización de lo vivo." *Systems Research and Behavioral Science*, *28*, 601–617. Original work published 1994. https://doi.org/10.1002/sres.1122.

Varela, F. J., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems*, *5*(4), 187–196.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.

Vázquez, M. J. C. (2020). A match made in heaven: Predictive approaches to (an unorthodox) sensorimotor enactivism. *Phenomenology and the Cognitive Sciences*, *19*(4), 653–684.

Venter, E. (2021). Toward an embodied, embedded predictive processing account. *Frontiers in Psychology*, *12*, 137.

Vernon, D. (2013). Interpreting Ashby–But which One? *Constructivist Foundations*, *9*(1), 111–113.

Vespignani, A., & Zapperi, S. (1998). How self-organized criticality works: A unified mean-field picture. *Physical Review E*, *57*(6), 6345.

Villalobos, M. (2013). Enactive cognitive science: Revisionism or revolution? *Adaptive Behavior*, *21*(3), 159–167.

Villalobos, M., & Dewhurst, J. (2018). Enactive autonomy in computational systems. *Synthese*, *195*(5), 1891–1908.

Villalobos, M., & Ward, D. (2015). Living systems: Autonomy, autopoiesis and enaction. *Philosophy & Technology*, *28*(2), 225–239.

Virgo, N., & Ikegami, T. (2013). Autocatalysis before enzymes: The emergence of prebiotic chain reactions. In *Advances in artificial life, ECAL*, *12,* 240–247.

Virgo, N., McGregor, S., & Ikegami, T. (2014, July). Self-organising autocatalysis. *ALIFE 14: The Fourteenth International Conference on the Synthesis and Simulation of Living Systems*. *14*, 498–505.

von Bertalanffy L (1968). *General system theory: Foundations development applications*. Penguin Books.

Walker, G. (2013). Evolution, co-evolution, and complexity: An anniversary systems journey through the Grid. *Civil Engineering and Environmental Systems*, *30*(3–4), 249–262.

Walmsley, J. (2008). Explanation in dynamical cognitive science. *Minds and Machines*, *18*(3), 331–348.

Ward, D. (2016). Hurley's transcendental enactivism. *Journal of Consciousness Studies*, *23*(5–6), 12–38.

Ward, D., Silverman, D., & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, *36*, 365–375.

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*(2), 97–125.

Weslake, B. (2006). Common causes and the direction of causation. *Minds and Machines*, *16*(3), 239–257.

Westheimer, G. (2008). Was Helmholtz a Bayesian? *Perception*, *37*(5), 642–650.

Whitehead, A. N. (1925). *Science and the modern world*. Macmillan.

Wicken, J. S. (1981). Causal explanations in classical and statistical thermodynamics. *Philosophy of Science*, *48*(1), 65–77.

Wiese, W., & Friston, K. J. (2021). Examining the continuity between life and mind: Is there a continuity between autopoietic intentionality and representationality? *Philosophies*, *6*(1), 18.

Windt, J. M. (2018). Predictive brains, dreaming selves, sleeping bodies: How the analysis of dream movement can inform a theory of self-and world-simulation in dreams. *Synthese*, *195*(6), 2577–2625.

Winter, O. C., & Murray, C. D. (1997). Resonance and chaos: I. First-order interior resonances. *Astronomy and Astrophysics*, *319* (1), 290–304.

Wittgenstein, L. (2010). *Philosophical investigations*. John Wiley & Sons. Original work published 1953.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Wostmann, B. S., Larkin, C., Moriarty, A., & Bruckner-Kardoss, E. (1983). Dietary intake, energy metabolism, and excretory losses of adult male germfree Wistar rats. *Laboratory Animal Science*, *33*(1), 46–50.

Wright, J. C. (2001). Cryptobiosis 300 years on from van Leuwenhoek: What have we learned about tardigrades? *Zoologischer Anzeiger-A Journal of Comparative Zoology*, *240*(3–4), 563–582.

Yoshimi, J. (2016). *Husserlian phenomenology: A unifying interpretation*. Springer.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zahavi, D. (2003). *Husserl's phenomenology*. Stanford University Press.

Zahavi, D. (2004). Phenomenology and the project of naturalization. *Phenomenology and the Cognitive Sciences*, *3*(4), 331–347.

Zhang (2008). Wittgenstein's reconsideration of the transcendental problem. *Frontiers of Philosophy in China*, *3*(1), 123–138.

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in Neural Systems*, *17*(4), 301–334.

Zhou, B., Bentham, J., Di Cesare, M., Bixby, H., Danaei, G., Cowan, M. J., & Cho, B. (2017). Worldwide trends in blood pressure from 1975 to 2015: A pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *The Lancet*, *389*(10064), 37–55.

Index