THERE AND BACK AGAIN

From Physics to Epistemic Agents

by

THIJS HEMME

First Supervisor:
DR. GUIDO BACCIAGALUPPI

Second Supervisor:
DR. MIGUEL SEGUNDO-ORTIN

Second Examiner:
PROF. DR. F.A. MULLER

Utrecht University

Contents

THERE AND BACK AGAIN

From Physics to Epistemic Agents

Thijs Hemme

I. INTRODUCTION

> It is true that the whole scientific inquiry starts from the familiar world and in the end it must return to the familiar world...
>
> —Arthur Eddington (1928, xv)

In his 1962 paper 'Philosophy and the Scientific Image of Man' Wilfred Sellars famously suggested that the goal of philosophy should be to know one's way around the subject matters of the various scientific disciplines as parts of the intellectual landscape as a whole, to build bridges between these scientific disciplines, and thereby "to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term" (Sellars 1962, 1). This manner of framing the enterprise of philosophy has become somewhat of a standard expression among philosophers of a naturalistic bend. Here I will follow suit. But for Sellars, the philosophical enterprise of mediating between our different modes of understanding is not just concerned with the body of provisionally accepted scientific theories—what Sellars dubbed the *scientific image*—but also with our more 'common-sense' view of the world—the *manifest image*. These two 'images', both of roughly the same order of complexity, were ideally to be reconciled into one 'stereoscopic view'.

Part of Sellars's inspiration for coining and introducing the notion of a 'scientific image' versus a 'manifest image' came from a well-known example due to Arthur Eddington. In the introduction to his 1927 *Gifford Lectures*, Eddington famously remarked how different his familiar understanding of a table was from his scientific understanding of a table:[1]

[1] Sellars (1962, 35-36) only briefly refers to Eddington's 'two tables' as an illustration of the two 'images'; Savitt (2012) and Callender (2017, 24-26) do so in more detail (in both cases with respect to issues pertaining to time, where they use the Sellarsian phrases 'manifest time' and 'scientific time'/'physical time').

> I have settled down to the task of writing these lectures and have drawn up my chairs to my two tables. Two tables! Yes; there are duplicates of every object about me—two tables, two chairs, two pens. …
>
> One of them has been familiar to me from earliest years. It is a commonplace object of that environment which I call the world How shall I describe it? It has extension; it is comparatively permanent; it is coloured; above all it is *substantial.* […] if you are a plain commonsense man, not too much worried with scientific scruples, you will be confident that you understand the nature of an ordinary table. …
>
> Table No. 2 is my scientific table. It is a more recent acquaintance and I do not feel so familiar with it. It does not belong to the world previously mentioned […]. It is part of a world which in more devious ways has forced itself on my attention. My scientific table is mostly emptiness. Sparsely scattered in that emptiness are numerous electric charges rushing about with great speed; but their combined bulk amounts to less than a billionth of the bulk of the table itself. (Eddington 1928, xi-xii)

Eddington admits that both tables appear to be perfectly accurate and workable descriptions of a table; he says:

> there seems to be nothing to choose between the two tables for ordinary purposes; but when abnormal circumstances befall, then my scientific table shows to advantage. If the house catches fire my scientific table will dissolve quite naturally into scientific smoke, whereas my familiar table undergoes a metamorphosis of its substantial nature which I can only regard as miraculous. (Eddington 1928, xii)

This is of course what primarily distinguishes the scientific image from the manifest image: the scientific image is more rigorous and precise, more general and less parochial, it is purportedly objective, whereas the manifest image results from our (shared) subjective perspectives. Eddington's two tables—the scientific table and the 'manifest table'—are a nice illustration of this 'dichotomy' between the two images. The two different views on a table are however, Eddington remarked, "two aspects or two interpretations of one and the same world", and ideally "they are ultimately to be identified after some fashion" (1928, xv). The challenge put forth here by Eddington is thus, in Sellars's words, how to formulate a stereoscopic view with respect to the two tables.

From this seemingly innocent little example, many philosophical and scientific controversies can be extracted regarding how one should go about doing something like this, if at all. One of the things that I think characterize a more 'naturalistic' approach to addressing this sort of problem is to *start with* the scientific table and try to *work back* to something resembling the manifest table—to go *from there to here* as Callender (2017) puts it. Thus, one could try to find an identification by beginning with the scientific table and to see if approximations to the manifest table can emerge at the scales at which beings like us operate (Callender 2017, 25-26). But physics alone will not be able to explain all the properties of the manifest table. If we want to understand properties like its color or texture, we will need to understand the human perceptual system and the interaction between the surface of the table and this system. We would thus need to embed a subject in the world and model its actions and abilities if we want to understand the manifest table. Al the while we can, and should, acknowledge that the science is ongoing. But armed with enough knowledge of physics and visual and tactile perception, we would be able to tell a convincing reconciliatory story that would make the manifest table a lot less surprising given the scientific table. Note that solving this 'two tables' problem is not at all easy, since much of the solid-state and condensed-matter physics, materials science, and cognitive science required is very much work-in-progress. Nevertheless, we generally assume that such a reconciliation *can* be found within our existing arsenal of scientific theories and models.

This 'two tables' example is relatively uncontroversial. Most would hopefully agree that the manifest properties of tables can be roughly recovered from their physical properties as described by physics in conjunction with a scientific description of the human perceptual system. However, it is precisely this kind of reconciliatory strategy that can be argued to be called for in many more—sometimes controversial—cases. Examples in the philosophy of mind abound where people insists that a purely scientific description of our world cannot possibly account for many of our manifest impressions of the world. I am inclined to disagree with this, but I do recognize that it takes a lot of work and many steps to flesh out an account in which something like it can be satisfactorily argued for.

Sellars's vision of philosophy as mediating between the scientific image and the manifest image and understanding how all things hang together may be somewhat ambitious. In fact, it would mean that to be a philosopher one is required to be a brilliant polymath, well-versed in extremely hard interdisciplinary work. Clearly that would be setting the bar a little high. That does not of course mean we should not aspire to it. This interdisciplinary conciliatory vision of philosophy can serve as a higher goal for philosophers, who may contribute to it piecemeal with specific case-studies. That is in part how I would like to think of the topic I discuss in this thesis.

This thesis is about how a certain combination of current scientific insights and ideas suggest how we may think of epistemic agents doing science as particular kinds of very complex physical systems. It is about how we could formulate an account where we go from a physical description of the world to one that contains epistemic agents that can devise that physical description in the first place. For that purpose I will draw on ideas emerging in complexity science and cognitive science. Much has been discussed regarding the relation between physics and the 'observer'. I intend for my discussion to also further clarify such issues by attempting to sketch a somewhat more complete story about how we might go from physics to epistemic agents.

This thesis will thus be focused around the question of how knowledge of the world, or science more specifically, is possible in a world that—on certain readings of our current scientific image—is 'ultimately' physical. There are two possible 'tensions' one could try to alleviate in that regard. On the one hand it concerns the more general issue of making sense of how a world 'governed by impersonal laws' could give rise to goal-directed and intentional phenomena. How do you go from pure physics to linguistic constructs by human agents that are taken to be 'theories' constructed for the *purpose* of *representing* the physical world? The issue here is more conceptual and 'metaphysical', since there seems a tension between the teleology and agency involved in human scientists and their products, and the purely mechanistic, dynamical physical laws on the other. That issue in itself is not wholly central nor unique to this thesis, as that is a much more general topic in the philosophy of mind and other areas. In another sense this tension can be identified as the issue of how relatively simple impersonal physical laws could possibly give rise to something as complex and delicately structured as intelligent epistemic agents constructing sophisticated theories and models that adequately represent those physical laws and phenomena emergent from them. This becomes especially puzzling given the second law of thermodynamics, which would naïvely lead one to expect that delicate structures tend to disintegrate, and that all systems would naturally tend to a turbulent, disordered state, and eventually thermal equilibrium. This thesis will primarily focus on this second tension, and although I will bracket the first tension here, I have the hope and expectation that a resolution to the second will suggest some resolutions regarding the first.[2]

In addressing how science is naturalistically possible, I will employ a naturalistic approach to philosophy, and through it I will try to formulate a naturalistic philosophy of science. A proper such naturalistic philosophy of science would have to incorporate insights from a great plurality of disciplines, only a few of which I shall attempt to do justice to here. In the following chapter I will first clarify what I mean by my naturalistic approach. Here I will discuss how my attitude aligns with the likes of Ladyman & Ross (2007), and how I would like to formulate a kind of naturalized epistemology, albeit slightly differently from how that is usually thought of. My approach is to start with a physical picture of the world—in accordance with Ladyman & Ross's principle of the 'primacy of physics'. Ideas about thinking of observers or agents as

[2] Cf. Deacon (2011).

physical subsystems have been formulated in other contexts, and the strategy utilized in those contexts, where certain distracting ‘metaphysical’ issues are bracketed, is one that I will employ here as well. However, the simple toy-models of observers appealed to in those cases, which do a perfectly fine job for the purposes at hand, will not suffice for my purposes. I will require a more detailed and realistic account that includes how these ‘observers’ come about. Using insights from emerging fields that study complex systems we can start to make sense of how physics could give rise to more complex and goal-directed behavior as seen in biological and cognitive systems (3.1). From there I will draw connections to current discussions in (the philosophy of) cognitive science, where I will try to incorporate the prominent new theory of ‘predictive processing’ (3.2). An especially interesting new idea that has arisen in these contexts is the ‘free energy principle’, a mathematical framework which proposes to describe what living and cognitive systems must physically do in order to persist and maintain their structure in a world where entropy persistently increases over time (3.3). This dynamical process can then be shown to be mathematically equivalent to the inferential processes as proposed in predictive processing. The free energy principle then purportedly explains how the apparent goal-directed behavior of living and cognitive systems can be thought of as a dynamical physical process. This might then allow for drawing a unifying connection between the realm of physics and the realms of life and cognition.

These ideas concerning complex systems, predictive processing, and the free energy principle provide a nice potential basis with which to address how science is naturalistically possible. The physical dynamics of complex systems far from equilibrium can give rise to behavior that through the lens of the free energy principle can be seen as *epistemic* behavior. We then still need to ascend several levels before we arrive at the human socio-cultural level of science being practiced in specifically designed communities and social institutions. Given that we have a plausible physical story to tell concerning how (apparent) epistemic behavior may arise from simple physical mechanisms, we can ask what more complex forms this is able to take in the case of humans. Using ideas concerning cultural evolution, evolutionary game theory (4.1) and the argumentative theory of reason (4.2), I think we can come a long way with reconciling how ultimately physical mechanisms can give rise to epistemic agents that engage in a cumulative discourse driven by norms of criticism and testability. We can furthermore make sense of that discourse construing and containing accurate representations of structures and patterns in the physical world by defining the real patterns in the world in terms of their parsimony, communicability and projectibility (4.3). Throughout, I will link some of my discussion to topics more traditionally discussed in the philosophy of science—

topics that I think I can give a more naturalistic gloss on. The end-product would then be a naturalistic philosophy of science.

The project here is thus to articulate how our scientific image might suggest that human epistemic agents can come about, engage in doing science, and to thereby construct that very scientific image. It is thus a somewhat recursive or 'holistic' view that I wish to articulate. The point here is not be provide a circular argument, but to provide a self-consistent view of science being possible in our scientific image. In articulating this naturalistic philosophy of science I will thus start at the level of physics and from there attempt to work my way up to our more 'manifest' understanding of scientists engaging in critical discourse driven by certain epistemic norms.

II. Epistemic Agents Naturalized[3]

Naturalism is a term that is used in many different contexts by many different people to denote many different things. My use of the term here will specifically refer to a certain style of thinking within philosophy of science. In that context 'naturalism' is still a fairly broad and vague term with no universally agreed-upon definition, where many philosophers and scientists often take it to involve some kind of combination of empiricism, pragmatism, and physicalism. This differs from how 'naturalism' in a broader philosophical context is often contrasted with a belief in the 'supernatural', and also sometimes even understood as something of a synonym for materialism or physicalism—or perhaps even scientific realism. These understandings of the term are somewhat misleading given the way I will use it here. Naturalism as a style of doing philosophy has more to do with a negative claim regarding the untenability of 'foundationalist' ideas in epistemology and the rejection of the notion of a 'first philosophy', and a positive claim that, as a result, philosophy needs to be 'continuous with' the natural sciences, and that philosophical questions can be addressed using scientific results and methods.

This idea that 'philosophy should be continuous with science' is however still a rather vague idea, and different naturalistic thinkers have spelled out that ideas in varied ways (Kincaid 2013, 2). There are nevertheless themes that naturalists generally agree upon and that roughly characterize the 'naturalistic stance'. This encompasses epistemological claims and metaphysical claims. The

[3] The first part of this chapter is adapted from an earlier (unpublished) paper titled *To Naturalize: Scientific Philosophy and Philosophy of Science* presented at the 2020 *Descartes Lectures—Pragmatic Naturalism: Progress in Science, Mathematics, and Values*, and submitted as a final paper for the 'Philosophy of Science Research Seminar' at Utrecht University as part of the History and Philosophy of Science research master's program.

epistemological claim is that all knowledge—including what we know about knowledge itself—comes from the application of the broad methods and results of the sciences; the metaphysical claim is, primarily, that any legitimate way of analyzing and studying what exists—of *doing* metaphysics—must be tied into the results and practices of sciences (Kincaid 2013). This naturalistic attitude does not come with a predetermined set of conclusions about what knowledge is and what the 'nature of reality' is; rather, what characterizes naturalism is the manner in which one reaches conclusions about those topics. The way naturalism is understood here is thus not in substantive terms of 'metaphysical naturalism' or 'epistemological naturalism' (e.g. Papineau 2020), but rather in terms of a *naturalized* metaphysics and a *naturalized* epistemology.

My understanding of naturalism resonates with that of James Ladyman and Don Ross, who with their 2007 book on naturalized metaphysics provided one of the strongest defenses of naturalism in the philosophy of science in recent years.[4] Ladyman & Ross's primary claim regarding a legitimate metaphysics is that philosophers need to defer to science, and physics especially, when doing metaphysics. To support this claim one needs to clarify what makes science epistemically superior such that one needs to defer to it in metaphysical matters. As opposed to the logical positivists' verifiability criterion of meaning, Ladyman & Ross propose that naturalists embrace a more pragmatist form of verificationism, where hypotheses that the 'scientific consensus' declares to be beyond our capacity to investigate should not be taken seriously (L&R, 29). 'Science' can here be understood as being demarcated from non-science solely by institutional norms—norms of scientific practice that are identified empirically. These norms do not have to be supposed to be arbitrary or the product of path-dependent historical factors. Science, as a community enterprise, achieves significant epistemological successes through collaboration and the creation of strong institutional filters for errors. Science, thus understood, in fact *is* our set of institutional filters for errors in the job of trying to construct empirically adequate theories or models that ostensibly describe the objective character of the world; and the epistemic supremacy of science rests on the repeated iterations of these institutional error filters (L&R, 28-29).

The naturalist can thus refer to institutional factors that make science epistemically superior, and use that to distinguish well-motivated from ill-motivated proposals, not to separate sense from nonsense. Instead of postulating direct epistemological criteria for determining what is deemed worthy of interest, institutional factors are taken as proxies, since the institutions of modern science are the most reliable epistemic filters we have—something that can be taken to have been established inductively by the specific

[4] Henceforth referred to as 'L&R'.

institutional processes of science over its history (L&R, 37). The 'naturalistic stance' that results from this is an explicitly normative one, and all the ideas and principles taken to be part of naturalism are to be considered as norms, as opposed to dogmas or doctrines.[5]

Under this understanding of the naturalistic stance, Ladyman & Ross argue that to naturalize metaphysics is to turn to science and explicate what the deep structures of our best theories appear to imply and claim about the nature of reality. With their proposal for a naturalistic metaphysics one of the important things that is desired from science is a relatively unified picture of the world. This is not asserted as a primitive norm; rather, unification is exemplified in the actual history of science, where scientists are often reluctant to pose or accept a hypothesis that is disconnected from the otherwise connected body of scientific hypotheses. An important feature for the justification of a hypothesis is "its standing in reciprocal explanatory relationships—networked consilience relationships—with other hypotheses" (L&R, 27). Metaphysics, of the naturalistic type, can then be understood as "critically elucidating consilience networks across the sciences" (L&R, 28).

In doing metaphysics naturalists should furthermore confer epistemic priority on physics over other sciences. A *metaphysical* hypothesis is taken to be one that unifies two specific scientific hypotheses, at least one of which stems from 'fundamental' physics. The motivation for endowing physics with this primacy comes from the history of science, where over the history of developments in physics physical forces have consistently been found, but no non-physical (chemical or 'living') forces ever have. Furthermore, scientists made progress in being able to unify and extend the physical forces that were found (L&R, 42). Physical hypotheses have thus been successful, unified, and extended, giving reasons for "supposing that there is a coherent body of fundamental physical theory of sufficient scope and power that it is the only candidate for the 'most basic and comprehensive of the sciences'" (L&R, 43). Note that when Ladyman & Ross speak of 'fundamental' physics they do not take themselves to be speculating about a putative physical 'bottom' to reality; instead they refer to 'fundamental' physics as that part of physical theories "about which measurements taken anywhere in the universe carry information" (L&R, 55), i.e. that are valid *everywhere* and at all scales.

In a similar spirit to Sellars, by unifying physical theories with other scientific hypotheses Ladyman & Ross take a *useful* metaphysics to be in the service of showing "how the separately developed and justified pieces of science (at a given time) can be fitted together to compose a unified world-view" (L&R, 45). It is good to stress again that for Ladyman & Ross this 'unity of science' is a working

[5] Ladyman & Ross take the notion of a 'stance' from van Fraassen (2002), which is taken to be a combinations of values, attitudes, commitments, forms of life, and so on.

hypothesis. An obvious issue with this naturalistic conception of metaphysics is that the metaphysical theory we may come up with is almost certainly going to be false, because the scientific theories on which it is based are likely going to turn out to be 'false'.[6] Scientific theories that are false can still be useful in guiding experiments or developing new theories, but it may be objected that it is not at all clear whether a false metaphysical theory is good for anything (L&R, 58). However, Ladyman & Ross argue that the 'weaker' form of metaphysics may be regarded as well-motivated, in contrast to 'strong' metaphysics. In this weaker form of metaphysics, philosophers doing metaphysics are actively involved in trying to resolve potential tensions between different well-tested and generally accepted scientific theories, and constructing a unified world-view in the process. In contrast to speculating about the true nature of reality 'from the armchair', such research may well help to progress science towards constructing an ever-more accurate picture of what the world may be like. This activity may be called 'metaphysics' because it does not have a specialized science of its own. Combined with a philosophy of science that supposes there is defensible basis for viewing the history of science as a history of progressive accumulation of knowledge, this naturalistic conception of metaphysics and this epistemic optimism reciprocally support each other.

Some critics of naturalism would argue that a naturalistic conception of ourselves threatens the very possibility of scientific knowledge, or metaphysical knowledge for that matter. Some would for example argue that it would supposedly not be possible to explain how natural selection makes scientific knowledge possible, since evolution does not optimize, it satisfies—evolution only provides us with a conception of the world as it is useful for us, not how it actually is.[7] This kind of objection holds little sway however, as closer inspections of evolutionary theory, cognitive science, and other sciences may very well yield an account of how scientific knowledge or (naturalistic) metaphysical knowledge might be possible (L&R, 6-7). Part of the theme of this thesis is to sketch precisely such an account.

This leads us into naturalized epistemology. Epistemology is the study of knowledge, and an accumulation of insights and critiques, especially in the last century, have culminated in the realization that any form of 'foundationalism' in epistemology is untenable. Insofar as there exists 'knowledge' of anything, it

[6] Perhaps a better way to put this is to say that currently accepted scientific theories (in physics at least) will probably turn out to be effective theories that only apply in certain regimes, but are not applicable across the board at all scales and in all regimes (i.e. 'fundamental').

[7] The evolutionary processes that gave shape to us are often assumed to have equipped us with reliable cognitive processes, since inferring things 'correctly' is generally more conducive to fitness than inferring things falsely. This could then provide some basis for a naturalistic epistemology (e.g. Quine 1969). That is of course an empirical hypothesis, and a plausible one, but some have questioned it. This premise has recently been questioned in a popular book by Donald Hoffman (2019) for example.

is not certain and not based on any firm unshakable foundation—everything we believe is *theory-laden* and thus relative to assumptions and some theoretical framework. This however need not lead to a strong form of relativism, for a variety of reasons.

In his 1969 paper 'Epistemology Naturalized' Quine famously and somewhat controversially suggested that epistemology should be naturalized. Often this is understood as the claim that the traditional philosophical questions of epistemology should make place for the study of scientific knowledge by cognitive science. Quine's arguments for this were twofold (Verhaegh 2018). First, all attempts at providing a foundation for knowledge had failed, such as Carnap's attempt at translating the sentences of science into terms of observation, logic, and set theory. But second, to naturalize epistemology is then not a move out of desperation, since all else failed; rather, Quine criticized the very project of providing a foundation for science as demonstrably flawed. This connects to Quine's famous critique of the analytic-synthetic distinction and his ideas concerning a holistic 'web of beliefs'. Instead, philosophy needs to 'work from within' (Verhaegh 2018) the framework provided by our best science when addressing philosophical questions, and he argued that it is "better to discover how science is in fact developed and learned than to fabricate a fictitious structure to a similar effect" (Quine 1969, 78). Note that this does not imply that all 'beliefs' are on a par; we can still recognize that certain constitutional principles play a very different epistemic role in our 'web of beliefs'—we can still distinguish between the 'hard core' and more peripheral 'auxiliary hypotheses'. But the difference here is not a difference in kind, but a difference in degree. And there can certainly still be useful philosophical work to be done in analyzing that 'hard core', such as the necessary preconditions for a theory and the basic tenets of a theory. But the shift in emphasis here is that one does not provide the foundations for science as such. Rather, one works with, and within, science.

The lesson I wish to take here for the epistemological dilemma could be characterized as approaching the question of knowledge about the natural world in a somewhat deflationary manner. Taking (some version of) the scientific image as our starting point, and as something that in our scientific discourse we take to provide us with a representation of a way the world might be, then within that way of talking about the world our scientific theories are trivially true. And within that scientific image, it must be the case that *we* exist and have come to know about those (trivially) true theories of the natural world, otherwise it would be self-undermining. The question we can then address is: how, within our scientific image, do human epistemic agents come about and come to acquire, or construct, knowledge of theories that accurately represent the natural world as stipulated by the scientific image. This circular, or self-consistent, way of studying what knowledge is, I think, is a good way to characterize what a naturalistic approach to epistemology could consist of. In fact, one of the things I think characterizes naturalism in the philosophy of science is the emphasis on a kind of 'virtuous circularity'. It is the emphasis on working from within the framework provided by our best scientific theories, instead of trying to provide *a priori* analyses from the outside. Theory-ladenness is seen as a feature, not a bug: embedded within a web of all our scientific knowledge, our philosophical concepts gain a much clearer meaning.

Another common objection to naturalism in epistemology is that it leaves no place for normative epistemic judgement since it replaces epistemology with 'purely descriptive science'. Now, first of all, it is a mistake to think of science as purely descriptive; scientists are essentially worried about what we are entitled to or ought to believe, and much of what they do is aimed at establishing evidence and reasoning about those obligations and entitlements (Kincaid 2013, 8). The reliability of methods, new and old, is a key scientific question toward which great attention and resources are directed. Naturalized approaches are not restricted to citing established facts about psychological processes in the way some people took Quine to suggest, and even those facts can have normative implications. This is not meant to imply that there is no crucial difference between descriptive and normative accounts. Normative accounts are concerned with competence, whereas descriptive account are about performance. Normative accounts of behavior generally involve creatures engaged in rule-guided communal activity, whereas descriptive accounts of behavior generally involve robust statistical correlations and causal relations between events. The question to address in order to reconcile these two accounts is then how it is that beings like us are able to engage in norm-governed interactions. What we need in the case of a naturalized epistemology is a kind of Sellarsian stereoscopic view, fusing our normative understanding of knowledge with a naturalistic understanding of how such knowledge could have come about. The 'descriptive' dimension should be about trying to provide a naturalistic framework that explains how epistemic agents are possible in the scientific image. This functions more like an 'existence proof' that a world such as suggested by the scientific image could give rise to epistemic agents like us. Such a naturalized epistemology is thus more concerned with naturalizing *epistemic agents*, and thus modeling them within our scientific theories. This thus fundamentally involves trying to reconcile our current normative understanding of what 'knowledge' is, with an account of how such knowledge could have come about 'naturalistically'.

This way of doing epistemology can incorporate many elements from scientific and philosophical work at many levels, and it will probably be an ever on-going project. As our scientific image keeps changing and improving, and as philosophical work on normative epistemological questions—such as work on formal epistemology, social epistemology, theories of explanation, etc.—keeps clarifying what we take justified knowledge claims to be, on-going work on a naturalistic epistemology will hopefully result in an increasingly improved understanding of what knowledge is and how it can be attained. This is one of the ways in which philosophy can be deemed continuous with science by the naturalist.

In (the philosophy of) physics a number of cases have come up that have in particular cried out for being clear and precise about modeling the observer in the theory.[8] Such an account of the observer turns out to be crucial for making sense of a whole slew of issues that come up and that some critics of these ideas put forth. In these contexts an approach has originated where the observer is modeled in the theory as an *information gathering and utilizing system* (IGUS) (Gell-Mann & Hartle 1990). The manner in which I will attempt to formulate a naturalistic epistemology can be thought of as similar in style to the way the notion of an IGUS has been applied to a variety of topics in (the philosophy of) physics. Central to this IGUS 'strategy' is to model an observer or agent in the physical theory as a physical subsystem that is capable of gathering and utilizing information about its environment. Bracketing more metaphysical issues surrounding 'aboutness' or 'consciousness', we can start with more operational questions concerning how we can recover, from basic physics, the behavioral manifestations and capacities that certain systems appear to display. As it keeps out a number of (initially) unnecessary distractions, I find this a fruitful strategy in the cases where it has been applied, and I believe it might be used in a similar way for a naturalistic epistemology.

In the context of knowledge about the natural world, Hartle (2016) has for example presented the 'anthropic' argument that, for the universe to contain IGUSes like human observers, it must exhibit regularities for these physical systems to exist, function, and proliferate. Up to some level the universe thus must be comprehensible: if there were no reliable regularities to exploit there would be no IGUSes. If a universe were so complicated as to be epistemically indiscernible, that universe would not have the requisite structure so as to produce living and cognizing systems to begin with. Thus, Hartle argues, given that IGUSes such as humans observers exist, there *must* be some regularities through which the universe can be comprehended. This is supposed to provide at least some basis for explaining why knowledge of the world is possible.

But an IGUS is often put forth as a rather simple toy-model of an observer or agent with some crude schematic structure—a rather simple computational device, capable of recording and storing information. This structure is just

[8] Two such cases are for example the four-dimensional spacetime, or 'block universe', of relativity and Everettian, 'many worlds', quantum mechanics. See e.g. Hartle (2005; 2016), Ismael (2015; 2017), Callender (2017), Saunders (1993; 1995).

postulated without any story as to how it could have come about or how it physically keeps itself in that configuration. In order to improve such a model-observer and make it more physically, biologically, and cognitively realistic we could use insights from contemporary complexity science and cognitive science. This then may inform us in more detail how such 'IGUSes' are capable of existing and persisting in the physical world and making inferences about their environments. This may thus result in a more realistic account of an IGUS as a physical subsystem that is genuinely describable as 'gathering' and 'utilizing' information from the physical world. This is what I will discuss in chapter 3. In chapter 4 I will furthermore discuss how a more sophisticated 'IGUS', or rather a community of 'IGUSes', could gain and construct more explicit knowledge of the physical world.

The purpose of this thesis is thus to illustrate how we might understand epistemic agents engaged in the normative discourse of science as ultimately physical systems, and their perceptual, cognitive, and behavioral capacities as ultimately physical processes. For this purpose I will draw on contemporary themes, ideas, results and discussions in a number of relevant scientific fields, as well as philosophical reflections on them. In my approach to this topic I shall try to incorporate elements from both naturalized metaphysics and naturalized epistemology.

III. A PHYSICS OF INFERENCE?

As a skeptic, David Hume introduced into philosophy some of the foundational problems with inductive and causal reasoning. As a pragmatist, however, he knew when to give it a rest. Setting aside the problems with rationally justifying to ourselves these forms of reasoning, Hume nonetheless recognized that causal and inductive reasoning does seem to 'fit' with the patterns of the world; and even if these forms of reasoning are ultimately the result of instinct, habit, or sentiment, it is hard to deny that these are good instincts and habits to have. While poking fun at Leibniz's idea of a 'pre-established harmony' between the purportedly causally disconnected constituents of the world (so-called 'monads') so as to explain the correspondence between our ideas and the world, Hume (1739, §V:44-45) suggests a resolution of sorts to the problem of induction:

> Here, then, is a kind of pre-established harmony between the course of nature and the succession of our ideas… As nature has taught us the use of our limbs, without giving us the knowledge of the muscles and nerves, by which they are actuated; so has she implanted in us an instinct, which carries forward the

> thought in a correspondent course to that which she has established among external objects.

Hume's 'naturalistic' solution to his own problem of induction is thus, as Harms (2004, 11), paraphrasing Hume, puts it: "nature *has* implanted in us cognitive instincts which keep our thoughts in productive harmony with the world, and this is the secret to understanding knowledge." Hume did not have an understanding of evolution, or how the physical dynamics of self-organizing processes give rise to complex systems. Nowadays we do—to some extent—and we can now understand better how exactly nature might make it so that our cognitive processes are in productive harmony with the world, and thus improve on Hume's 'solution'.

3.1 Self-Organization & Complex Adaptive Systems

How can the relatively simple laws of physics give rise to complex structures, such as living creatures? Naïvely one might think that the second law of thermodynamics is incompatible with complexity arising and increasing in the universe. There are two quick things to point out there: (i) the Earth is not a closed system, and (ii) complexity is not the same thing as low entropy. The second law applies to closed systems; in *open* systems, that exchange energy and information with other coupled systems, entropy *can* do down. Since the Earth is an open system, one that receives energy from the sun and radiates into the universe, complex structures forming on Earth is completely compatible with the second law of thermodynamics. Although that defuses some concerns, it does not address the real issue. This may explain why organized low-entropy systems *can* come into being here on Earth, it does not explain why they actually *do*. Furthermore, merely lowering the entropy of a system does not mean that complex structures arise. For example, we can locally lower the entropy of the contents in a refrigerator, but that does not make those contents more complex. We then still need to understand how complexity arises, and how and why the laws of physics bring it about.

It has been suggested that, as systems evolve from low entropy to high entropy, complexity is allowed to form in the intermediate stages.[9] Both low and high entropy states are 'simple' and non-complex, in the sense that they can be described with relatively few parameters and that no intricate structures are present. In the process of increasing entropy in a system, complex structures can come into being. Aaronson et al. (2014) have put forth the analogy of mixing cream into coffee, where only in the intermediate stages there are complex

[9] See e.g. Aaronson et al. (2014) and Carroll (2016, 225-313).

fractal-like tendrils of cream reaching into the coffee in intricate ways. The right way to think about complexity may thus be that, as entropy goes up, complexity first goes up and then goes down, in an inverse U-shape fashion. We can take this coffee and cream story to be a direct analogy to our universe. The universe 'started out' with low entropy initial conditions, and will likely evolve towards high entropy thermal equilibrium in the far future. Today, when the universe is medium-entropy, there is a lot of complexity.

Aaronson et al. (2014) have tried to make this idea more quantitative with simple simulations of 'cream' mixing into 'coffee'. As a proxy for complexity they measured the incompressibility of the images of these simulations of mixing cream and coffee 'particles'.[10] They found that, in a measurable sense, the 'complexity' of these images does indeed tend to take an inverse U-shape over time. With the appropriate combinations of long- and short-range interactions between the particles, the boundaries between the cream and coffee tend to take on fractal-like shapes.[11] Our actual universe also has both short-range forces (weak and strong force) and long-range forces (gravity and electromagnetism). What thus may be occurring when complex structures arise in our universe as entropy increases is an interplay between competing forces pushing and pulling on matter. The crucial thing to recognize here is that the appearance of complexity is not just compatible with increasing entropy, but that it relies on it. In a high-entropy universe complexity would never develop (apart from rare random fluctuations). The only reason complex structures form is because the universe is undergoing a gradual evolution from low entropy to high entropy. The growth of entropy is precisely what permits complexity to appear and endure. The 'cream and coffee' analogy may thus provide us with a simple intuition pump for understanding how all the complex phenomena we see around us can emerge out of basic physical rules, and not only 'despite' the second law, but because of it.

In 1790 Kant famously stated that there would never be a 'Newton of a blade of grass'. Some have subsequently argued that, in fact, some seventy years later this Newton of a blade of grass did come along, namely Charles Darwin. Arguably, however, Darwin was not really this proverbial Newton of biology (Schuster 2011; Deacon 2011). The theory of evolution by natural selection follows from the assumption that viable living systems that can replicate and on

[10] Do note that complexity understood in terms of ordered behavior and delicately organized structures is not the same thing as being 'complicated', which is essentially what is measured with incompressibility, also called 'computational complexity'. Although there must certainly be some connection between complexity and computational complexity, exactly what this connection might be is an ongoing topic of research.

[11] Fractality is a real indication of complexity, since fractals are geometric structures where intricate patterns manifest at all scales.

which selection can act are already in place. As Darwin himself eloquently phrased it in the last sentence of his *Origins of Species*:

> There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that…, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved.[12]

But exactly how life is 'breathed into' those forms was of course the question that Kant was after. Darwin's theory did not aim to address this. But in recent decades we have been gaining a better understanding of the self-organizing physics of living systems.

In 1944 Erwin Schrödinger famously speculated about the physics and chemistry required for life in his book *What Is Life?*. Schrödinger asked himself: "How can the events *in space and time* which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" (Schrödinger 1944, 3). To address this Schrödinger raised two issues that were to be addressed by the scientists of his time. The first question he posed was related to the 'hereditary mechanism'—how information was stored and transferred—and his speculations about this requiring an 'aperiodic crystal' ultimately inspired and led to the discovery of the double-helix structure of DNA. The second question, however, was related to how living systems are capable of maintaining themselves in a relatively low entropy configuration in a world where the second law of thermodynamics reigns supreme. As Schrödinger put it, the characteristic feature of a living system is that "it goes on 'doing something', moving, exchanging material with its environment, and so forth, and that for a much longer period than we would expect an inanimate piece of matter to 'keep going' under similar circumstances" (1944, 69). To be able to do this, Schrödinger argued, a living system must be using sources of thermodynamic free energy (or what he called negative entropy, or 'negentropy') in its environment to perform whatever physical work is required to maintain its structural integrity. But exactly how living systems are able to do this, one could argue, has still not been fully satisfactorily addressed.

Schrödinger's first question concerning the storage, transfer and replication of information arguably has been fairly satisfactorily accounted for by modern molecular biology. And all the molecular structures and mechanisms that have been uncovered in modern biology can be acted upon by Darwinian evolutionary processes—setting aside novel, and still to be better-understood, insights in evo-devo, epigenetics, and the like. However, this 'modern synthesis' of molecular biology and Darwinian evolutionary theory can be argued to only really address

[12] From the first edition (1859).

Schrödinger's first question, and thus not give us a full understanding of how living processes fit into the paradigm of physics and chemistry.

The distinction between Schrödinger's two questions can also be seen to be mirrored in the contemporary discussions about the origin of life, where opinions are roughly divided between the 'replication-first' camp and the 'metabolism-first' camp. Interestingly, whereas molecular biologists tend to side with the replication-first side of the debate (e.g. RNA-world hypotheses), people in complex systems research seem inclined to side with the 'metabolism-first' side (e.g. Kauffman 2019). Perhaps unsurprisingly, as much research in the field of complex systems nowadays focuses somewhat more on Schrödinger's less well understood second question, that is: on the dynamics of self-organizing processes.

The scientific study of complex phenomena came about in the second half of the twentieth century.[13] Complex systems are systems with a great number of moving and interacting parts that display ordered behavior—behavior that one would not easily be able to infer from just knowing the constituent parts. In recent decades ever more has been understood about the self-organizing dynamics of complex systems. Many of the insights gained in this field are highly technical and involve advanced mathematical techniques from non-linear dynamics and computer simulations. Nowadays the study of complex systems is a booming field, with some touting the science of complexity as *the* science of the twenty-first century (West 2017). The field is however still in a somewhat pre-paradigmatic state. There are no universally accepted theoretical frameworks and example cases one can point to that adequately capture all or a broad class of complex phenomena, and often researchers thus address very domain-specific cases. Furthermore, it is not clear whether there even is a single measure or definition of complexity to be found that captures what is meant with 'complexity' in complex systems (Mitchell 2009; Ladyman & Wiesner 2020). That does not of course mean proposals for these sorts of frameworks are not occasionally put forth. Much has been written about topics like self-organization, complex systems, synergetics, criticality, universality, emergence, and a handful of other related terms. I will not try to present a representative overview of all that has been said in that regard here. The main point I wish to convey is that contemporary insights are accumulating concerning these topics, which have demonstrated that, under the right circumstances, complex ordered behavior can arise out of relatively simple physical mechanisms in systems far from equilibrium.

[13] Aspects of the history of advances in the field of complex systems in the latter half of the twentieth century are wonderfully recounted in popular science books such as Gleick's *Chaos* (1987), Waldrop's *Complexity* (1992), and Strogatz's *Sync* (2003).

Following Deacon (2011), we can start to get a conceptual handle on understanding complex systems by making the distinction between self-organization and complex adaptive systems.[14] Self-organization is the physical process of the formation of patterns and structures in open systems far from thermodynamic equilibrium. Early work on non-equilibrium statistical mechanics by people like Ilya Prigogine, in his study of 'dissipative structures', and Hermann Haken, in his study of 'synergetics', introduced a more quantitative understanding of this process of self-organization. In this process some form of overall order arises out of the purely local interactions between the parts of an initially disordered system. The patterns and structures that emerge out of this process are formed through a variety of feedback mechanisms, which can keep going as long as enough free energy is available, and are generally robust against external perturbations. Self-organizing processes can form complex systems at a variety of scales. As Deacon (2011) argues, self-organizing systems can self-organize into a larger system in a dynamically coupled and hierarchically nested fashion. Complex *adaptive* systems are a special class of complex systems that can arise out of this hierarchical scaffolding of self-organization. These systems are adaptive in the sense that they have the capacity to change their state depending on the environment and can 'learn' from their environment.

An obvious example of complex adaptive systems are living systems. In line with a complex systems perspective on life, many contemporary ideas in the physics of life, like e.g. Jeremy England's (2013; 2020) theory of 'dissipation-driven adaption', are hinting at a fascinating connection between non-equilibrium thermodynamics and the self-organizing dynamics of living systems. The general picture that tends to emerge out of this complex systems view on life is that, physically speaking, life appears to be a 'solution' the universe stumbled upon to liberate sources of free energy and convert it into disordered high-entropy energy—in accordance with the second law of thermodynamics. Phrases have been uttered like life being 'nothing but an electron looking for a place to rest', or the purpose of life being 'to hydrogenate carbon-dioxide'. Emerging from these lines of research in the physics of life is thus the idea that life may in fact be a rather generic consequence of the laws of physics.

In section 3.3 I will discuss the 'free energy principle' as a proposed general framework describing the statistical characteristics that complex adaptive

[14] In his theory of 'emergent dynamics' Deacon (2011) proposes a generalization of thermodynamics by putting forth the idea that thermodynamic systems can be divided into three general categories, namely classic thermodynamic systems, self-organizing systems, and complex adaptive systems, for which he coins the terminology 'homeodynamics', 'morphodynamics', and 'teleodynamics'. I will be using more conventional terminology here.

systems, especially living systems, must have in virtue of their existence. The intriguing possibility that comes out of this framework is that the proposed dynamical process can be linked to quantities in information-theory and Bayesian statistics, and can thus be understood as an implicit inferential process. This suggests a continuity between life and cognition, and might explain how both can arise out of basic physical mechanisms. But before I get to the free energy principle, I will first say a few things about the context in which this idea arose, a context that will also be relevant for understanding human cognitive agents as will be further discussed in chapter 3.

3.2 Dynamical Agents & Predictive Processing

Cognitive systems are immensely complex systems. In fact, the human brain is often characterized as the most complex structure in the universe that we know of. But if cognitive systems are ultimately just very complex physical systems, we may ask how cognition arises out of ultimately physical processes. Ever since the invention of the computer, an existence proof had been provided to show that purely physical processes can behave ‘intelligently’, or at least that they can process information, perform logical operations, carry out computations, and solve problems. Thus, many philosophers and scientists hypothesized that cognition, and specifically the human mind, might be some sort of computational process implemented by the brain. This ‘computational theory of mind’ understandably became a widely embraced position in the cognitive sciences and the philosophy of mind, but in recent decades many authors have pointed out problems with this framework.

There are roughly three angles from which the computational framework has in recent decades been criticized; that is, by emphasizing the situated, embodied, and dynamical aspects of cognition.[15] These critiques all have in common that they focus more on concrete action, and emphasize that the way in which a cognitive agent’s behavior arises is through the dynamical interaction between its brain, body, and environment (Beer 2014, 128).

The situatedness, or embeddedness, of a cognitive agent concerns the interaction of an agent with the environment. While computational approaches tend to focus on abstract reasoning, situated approaches emphasize that the ultimate job of a cognitive system is to *do* something in its environment, to take concrete actions that have real consequences beyond its skull (Beer 2014, 129). Beyond that basic fact, the immediate environment also plays a central role in the behavior of an agent by providing a rich source of constraints and

[15] Here I primarily draw on an overview of these approaches as provided by Randall Beer (2014).

opportunities, as well as providing a context that provides meaning to the agent's possible actions. The relationship between the agent and the environment is furthermore one of ongoing interaction, where the environment is not just some source of isolated problems to solve. Rather the environment is a partner which the agent is engaged with in a continual improvisational dance. In situated approaches to cognition abstract reasoning is thus relegated to more of a supporting role, one that should be seen as a recent evolutionary elaboration on a more basic capacity for getting around in the world.

The embodiment of a cognitive agent concerns the fact that a nervous system is always part of a body, and that a cognitive agent uses its body to perform actions in its environment. The particular physical aspects of an agent's body crucially shape its behavior, and embodied approaches emphasize the way in which the body mediates all the physical interactions a cognitive agents engages in (Beer 2014, 132). Embodied approaches stress that cognition should to a large extent be understood as something biological. Therefore the biological aspects of an organism should also be taken into account, such as its evolutionary history, physiology, development, and the relevant neurophysiology. Furthermore, although human agents are capable or more abstract reasoning, embodied approaches tend to emphasize that even there our most abstract concepts are ultimately grounded in bodily experiences and body-oriented metaphors. The general thought here is thus that truly intelligent behavior requires a real body in a real environment, that the body gives shape to possible behavior, and that the biological features of organisms matter significantly to their behavior and cognition. A stronger claim that is also sometimes made is that the biological processes of living are fundamental and indispensable to cognitive capabilities, and that cognition is a basically an extension of a more basic imperative of living systems' capacity for autopoiesis and homeostasis.

Concerning the dynamical aspects of cognitive agents, what is often emphasized is that cognition should perhaps not be understood as a computational process. Rather, cognitive agents should perhaps be understood as dynamical systems, and cognition as the state-space evolution in these systems. Dynamical systems theory is a general mathematical theory that can be used to model how the state of a system changes over time in some systematic way. Applied to cognitive science, the dynamical perspective provides a different set of concepts, intuitions, and metaphors for thinking about cognitive systems. A cognitive system would them be a set of variables whose values evolve concurrently over time. In the dynamical systems framework the importance of time, context, interaction, embodiment and the environment are typically emphasized, and it is thus a natural ally of more situated and embodied approaches to cognition (Arkoudas & Bringsjord 2014, 54). This dynamical systems approach has also proved fruitful in many areas of cognitive science, and it is now an essential tool in computational neuroscience (Beer 2014, 135).

Another interesting point raised by van Gelder & Port (1995, 28-32) is that digital computers are in fact also dynamical systems at the level of electrical circuitry, but these systems have been deliberately constructed so that they are amenable to a coarse-grained description of them as discrete computational systems at a higher level. The computational approach might then be rephrased as the hope that evolutionary processes have shaped cognitive processes and the nervous system so that they are also amenable to a computational description. Dynamicists tend to find this an unrealistic expectation for complex systems that evolved for their behavioral efficacy and not for their intelligibility in terms of engineering design principles. At most, only certain cognitive processes might be amenable to a computational description, or it might provide an approximation to them. Dynamical approaches moreover still allow for information-processing to be part of cognition, but only insofar this can be understood as a dynamical process, as for example in neural networks.

Combining these situated, embodied, and dynamical perspectives on cognition, an alternative to the computational way of thinking about cognition can be formulated—what Beer (2014) calls the 'SED' framework. First, we can think of brains, bodies, and environments as coupled dynamical systems that are each characterized by their own set of states and whose temporal evolution is governed by dynamical laws. As coupled systems, brains are embodied in bodies, and bodies are situated in environments. The coupled brain-body system can be thought of as the 'agent', where couplings flowing from the environment to the agent can be seen as sensory, and couplings flowing from the agent to the environment as motor (see fig 3.1). The sensorimotor cycles in this coupled brain-body-environment system are a property of the whole system and the behavior thus cannot be singularly attributed to one subsystem in isolation from the others. The proper object of study under this perspective is thus the full trajectory of the complete brain-body-environment system. It can however still be meaningful to ask what the relative contributions of the brain, body, or environment are to some particular feature of a behavioral trajectory (Beer 2014, 138). But to think in terms of a 'disembodied mind' would, under this perspective, be an oxymoron—the thought experiment of a 'brain in vat' would, even if possible in principle, be a completely misguided way to understand cognition.[16]

Similar approaches to this SED framework also sometimes go under the name of '3E' (embodied, embedded, and enacted) or '4E' (+ extended) approaches to

[16] Furthermore, dynamicists might argue that the 'cognition' with a brain in a vat happens in the dynamical interaction of the brain with its simulated body and environment.

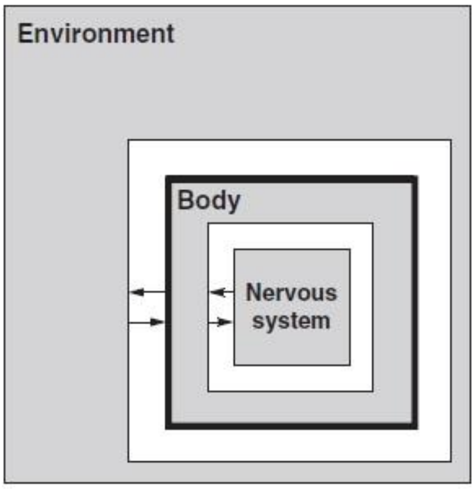

Fig. 3.1: A diagram depicting the brain, body, and environment as coupled dynamical systems (from Beer (2014, 137)).

cognitive science.[17] There are certainly differences between the separate components of these acronyms, and different authors taking these approaches can differ in their ideas significantly. Nevertheless, the general idea behind these approaches is that brain should not simply be thought of as a computer—it is not running a piece of 'mind' software on the hardware of the brain or nervous system. Instead, what is emphasized is that we should take account of how cognition is to a large extent biological, social, and dynamical.

Some have also noted some similarities between this development in cognitive science and other intellectual disciplines (Arkoudas & Bringsjord 2014, 57). In the philosophy of language, for example, there has been a shift of attention from 'the sentence' as an abstract theoretical entity to concrete speech *acts* carried out be real people in real time. Similarly, in philosophy of science, it is more often stressed that science is a human *activity* that is contingent on social interaction and cultural and political factors. In general, as Arkoudas & Bringsjord (2014, 57) note, one can see that "there has been an overall trend away from statics and toward dynamics, from the abstract and decontextualized to the concrete and context-bound, from justification to discovery, from isolated contemplation to social interaction, and from thinking to doing". The recurring theme that keeps gaining prominence here is that something which is dynamic, evolving, reactive, plastic, flexible, informal, highly nuanced, textured, colorful, and open-ended can never by properly understood by modeling it as something which is static, rigorous, unbending, and inflexible.

In recent years Bayesian accounts of brain function have gained in prominence in neuroscience. Particular versions of these accounts have been put

[17] 'Enacted' refers to the idea that perception and cognition are closely connected to action, and cognition is thus something 'enactive'; 'extended' refers to the idea that cognitive processes should not be seen as limited to the brain and body, but that the environment should also be recognized as a part of cognitive processes.

forth as being able to reconcile the more 'traditional' computational approaches to cognition with the more situated, embodied, and dynamical approaches. Bayesian accounts of brain function propose that the brain can be seen as an inference engine that operates in situations of uncertainty in a manner close to optimal as prescribed by Bayesian statistics. In these accounts it is assumed that the nervous system maintains internal probabilistic models, and that through neural processes these models are updated in order to minimize prediction errors. One increasingly prominent such theory in cognitive science and neuroscience is 'predictive processing' (PP). The general theoretical framework of PP proposes a unifying account of the workings of perception, cognition, and action.[18] As a theory of perception, PP states that "perception is the result of the brain inferring the most likely causes of its sensory input by minimizing the difference between actual sensory signals and the signals expected on the basis of continuously updated predictive models" (Seth 2015, 1). For this reason perception has also been characterized as a kind of 'controlled hallucination', since what is 'perceived' is the predictive model, not the actual sensory signals, with the perceptual system continually in the business of trying to minimize the difference between the two.[19] As a theory of action, PP states that, instead of updating the predictive models, the brain can also 'steer' the body in order to minimize prediction errors. From these two basic mechanisms, PP purports to be able to account for all of perception, cognition, and action, including when they malfunction (such as with perceptual illusions and psychological pathologies), and the interplay of these two mechanisms at multiple scales is thought to give rise to the whole complex behavioral repertoire of animals.

Biologically speaking, the function of brains and nervous systems is fundamentally to control the body—both to regulate the internal states of the body and to guide the behavior of the body in the environment in an adaptive way. But the brain does not have any direct causal access to the body and the environment. What it does have access to is the information in the flux of incoming sensory signals. This information, however, is argued to be noisy and ambiguous. According to PP, the brain has an 'expectation' of what the continually incoming sense data are going to be. What the brain then has access to is the discrepancies between these expectations and the sense data, and the brain can either 'steer the vessel' or change the expectations so as to minimize these discrepancies, and as such is capable of producing adaptive behavior. In

[18] For a comprehensive introduction to predictive processing, see e.g. Jakob Hohwy's *The Predictive Mind* (2013) and Andy Clark's *Surfing Uncertainty* (2016).

[19] This perspective on perception also connects nicely to the notion of the theory-ladenness of observation. In fact, thinkers such as Hanson and Wittgenstein who put forth these ideas in the twentieth century made extensive use of aspects of our perception and ideas from gestalt psychology that we can nowadays neatly explain in terms of predictive processing.

terms of brain physiology, PP specifically proposes that the hierarchical structure in the brain is such that different levels can feed their in- and outputs to each other according to particular precision weighting mechanisms, and that through this cascade of upward and downward signals the whole complex behavioral repertoire of animals arises. On this account of perception, instead of a passive recipient of percepts, the brain and body are active and anticipatory participants in the process of perception. There is a fair amount of psychological phenomena that this idea can account for, leading to PP nowadays being a main contender for a theory of brain function.[20]

The idea that perception has something to do with predictive control is very old, going back to cybernetics, Helmholtz, and arguably even Kant. In the nineteenth century the German polymath Hermann von Helmholtz first formalized the idea of perception as 'unconscious inference' (*unbewusster Schluss*). Helmholtz formulated the idea in part to account for a whole slew of visual illusions; as he phrased it: "Objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism."[21] The Helmholtzian view, however, was a rather passive one, with no connection to action and behavior. These ideas were also long ignored in psychology and philosophy, and only really received renewed attention in the latter half of the twentieth century. In the twentieth century, predictive coding schemes, first developed in computer science in order to compress file sizes, found their application in neuroscience as a formal account of how these unconscious predictions might be realized in the brain. Implementations of this kind in machine learning were also dubbed the 'Helmholtz machine' by Dayan et al. (1995), which they intended as a model for the human perceptual system.[22] All these ideas, it may be said, later synthesized into the more passive component of what is now called PP.

The closer link between perception and action could be seen in the work of the mid-twentieth-century cyberneticians such as W. Ross Ashby (1957; 1960). This could especially be seen in Ashby's work on the 'homeostat' and the 'good regulator theorem'. Ashby emphasized the importance for adaptive systems of the homeostasis of internal essential variables, that is, of regulating these essential variables to stay within viable bounds. This, Ashby proposed, could be achieved with a variety of positive and negative feedback mechanisms. This cybernetic principle thus proposes that the general purpose of adaptive (e.g. biological and cognitive) processes lies in maintaining homeostasis to ensure

[20] See e.g. Seth (2015; 2019; 2021); for the empirical evidence that supports PP, see de Lange et al. (2018) and Walsh et al. (2020).

[21] From Helmholtz (1925), an English translation of Helmholtz (1879).

[22] As Dayan et al. (1995, 889) put it succinctly: "Following Helmholtz, we view the human perceptual system as a statistical inference engine whose function is to infer the probable causes of sensory input".

that internal essential variables remain within expected ranges, and "that adaptive systems ensure their continued existence by successfully responding to environmental perturbations so as to maintain their internal organization" (Seth 2015, 8). In order for a successful control system to achieve this feat, Ashby argued, it must have a good model of the system to be controlled. He formalized this in the 'good regulator theorem' (Conant & Ashby 1970), which states that *every good regulator of a system must be a model of that system* in the sense that there must be a homomorphic mapping between the regulator and the system to be regulated. This means that a successful control system must be capable of entering at least as many states as the system being controlled. In this cybernetic picture we thus get a picture of adaptive processes as *actively* in the business of regulating and achieving homeostasis, which it can achieve by instantiating a model of the variables to be controlled.

Many of the ideas in PP draw on these older cybernetic ideas. The modern idea is that prediction error can be minimized in two ways: by changing the predictive models, called 'perceptual inference', or by taking actions to change the body or environment and thereby confirm or test sensory predictions, called 'active inference'.[23] Here, this active inference form of PP could also be viewed as a more formal account of embodied cognition, and PP more generally is eminently compatible with situated and embodied approaches to cognitive science (Seth 2015; 2019; Clark 2015a; 2015b). Active inference is the context in which Friston originally devised the mathematical machinery of the free energy principle. Predictive processing has subsequently been characterized as an 'implementation' or a 'process theory' of the more general, first-principles framework proposed by the free energy principle.

3.3 Bayesian Mechanics

The free energy principle (FEP) began initially as a mathematically principled way of formulating PP in neuroscience. The information-theoretic 'free energy' quantity was seen to be a calculably tractable measure of prediction error that the brain could minimize. The FEP thus began as an account of the dynamics and function of the nervous system. Thereafter Friston and others started speculating about applying the FEP to livings systems more generally, where the FEP might provide a principled account of the dynamics of all living systems and subsystems, as well as the behavioral dynamics of groups of living systems (Ramstead et al. 2018; Hesp et al. 2019). With such ambitious universal claims, the FEP has in recent years attracted its fair share of supporters and critics, and

[23] Note that sometimes the label 'active inference' is also used to denote the predictive processing theory as a whole (e.g. in Parr et al. (2022)).

in the literature considerable confusion exists concerning its meaning, applicability, and merits. Friston has characterized the FEP as a principled account of 'sentient behavior' grounded in physics. He and others sometimes consider the FEP to provide something of a 'physics of life and mind', and more specifically the FEP is claimed to be a variational principle of least action for livings systems. These are still speculative ideas that are contested, but if the FEP and its purported consequences are indeed legitimate, then it might hold great potential for unifying ideas in physics, biology, and cognitive science, as well as providing something of a physical and biological foundation for a naturalistic epistemology.

The rationale for the FEP in its current formulation is that it provides a first-principles account of self-organizing adaptive behavior by starting from extremely general considerations in a similar spirit to the cybernetic tradition: what would a system physically have to do in order to maintain its structural integrity and to persist for a non-trivial amount of time, and thus locally resist the increase of entropy? In order to address this question Friston drew on older ideas from cybernetics concerning homeostasis and the 'good regulator theorem', and combined a collection of mathematical techniques stemming from statistical physics, dynamical systems theory, machine learning, and information theory to come up with what is now the FEP.

'Free energy' in the FEP refers to the information-theoretic quantity also called 'variational free energy', not the various forms of free energy as they are known in physics. The terminology stems from the variational method of free energy minimization initially developed in statistical physics in the 1970s for the purpose of approximating computationally intractable quantities. This same variational method of free energy minimization was later adopted in statistics and machine learning for the purpose of approximating Bayesian inference when intractable integrals show up, as they most often do for evaluating the marginal likelihood. The 'variational free energy' quantity that is minimized in these contexts is thus a purely formal quantity without physical units. Friston took this free energy minimization technique and re-applied it to ideas in statistical physics and dynamical systems theory (see the Appendix for a sketch of the formalism involved).

The FEP builds up from a fairly trivial starting point. It starts by writing out the rate of change of the probability density function of a 'random dynamical system'—a system whose temporal evolution is governed by a deterministic flow plus random fluctuations—using the 'Fokker-Planck equation'.[24] Such a random dynamical system is taken as a model of how natural systems evolve in accordance with the second law of thermodynamics. As per the second law, the

[24] A probability density function is a continuous distribution that describes the relative probability of finding a system in a certain state at a particular time.

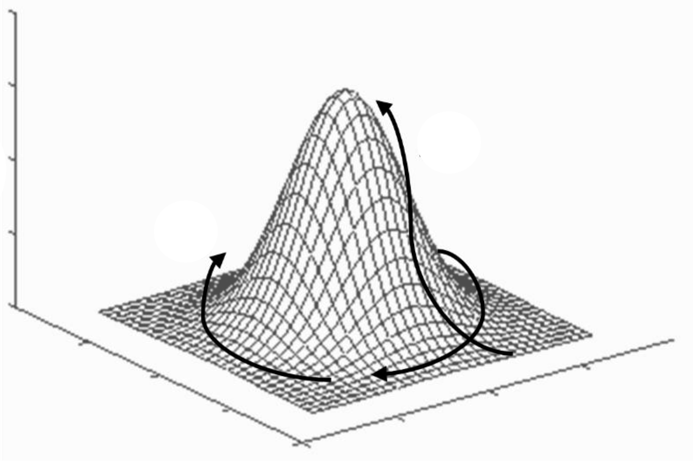

Fig. 3.2: Gradient ascent on a probability density function.

random fluctuations would ordinarily cause the probability density function to spread out over time, thus increasing the system's entropy. Since we are interested in a system that maintains its structure over time, this equation is set equal to zero, which means that the probability density function remains constant over time and the system is most likely to stay within a fixed region of its configuration space. Such a static form is also known as a 'non-equilibrium steady state' (NESS), in which there are non-zero flows in the system, but on average there is no time variation in the variables defining the system. In such a static form, the Fokker-Planck equation can be solved for the deterministic flow of the system. This solution can be parsed, in what is called a 'Helmholtz decomposition', into a curl-free and a divergence-free flow component that both evolve as a function of the logarithm of the static probability density of the system. The curl-free, or irrotational, flow component counters the random fluctuations, whereas the divergence-free, or solenoidal, flow component as a circular current leaves the system within the same region of its configuration space. Viewed as flows on the landscape of the NESS-density, the irrotational flow can be seen as performing a gradient ascent on the logarithm of this density, whereas the solenoidal flow circles around the contours of the logarithm of the density (see fig 3.2). As such, the irrotational flow counters the dispersion of the density and the solenoidal flow leaves the density the same everywhere.

The statistical quantities at play here can also be interpreted through an information-theoretic lens. The negative logarithm of a probability distribution is also known as the 'Shannon information' or 'surprisal', which is a measure of the 'unexpectedness' of some outcome or signal given some probability function. This thus also allows us to say that the irrotational flow performs a gradient *descent* on surprisal, and as such makes it so that the system stays within less 'surprising' regions of its configuration space given the NESS-density. Furthermore, minimizing surprisal (or maximizing the logarithm of a probability function) is equivalent to maximizing marginal likelihood, which in

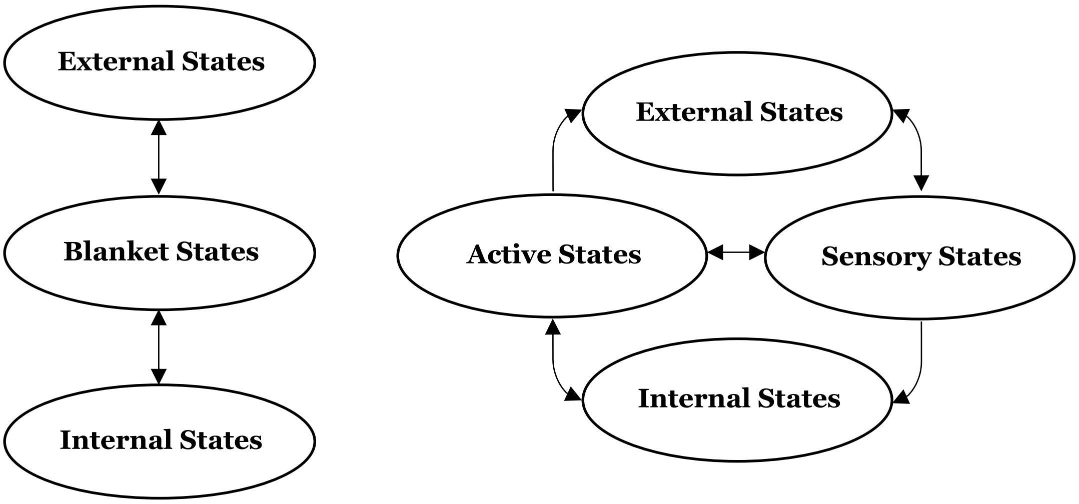

Fig. 3.3: A causal Bayesian graph depiction of a Markov blanket (left), and Friston's separation of the blanket states into active and sensory states (right).

the context of Bayesian statistics is how one optimizes the evidence for a 'belief', and thus optimally performs Bayesian inference given some prior belief. This becomes relevant once a partition is introduced in which part of the system can be interpreted as performing Bayesian inference with respect to itself and the rest of the system.

The next step thus concerns partitioning the system that is in a non-equilibrium steady state into states that are conditionally independent from each other. As such, one can identify a subsystem that is meaningfully separate from the rest of the system, but still causally connected to the rest of the system. This is intended as a model of measurable identifiable *open* system and the environment it exchanges matter and energy with, and thus interacts with. Here Friston makes use of what is called a 'Markov blanket', an idea originating in Pearl's (1988) work on causal modeling. A system with a Markov blanket is a system in which one can identify *internal* states that are conditionally independent from *external* states due to the presence of *blanket* states (or a Markov blanket) that 'shield' the internal states from the external states by mediating influences. Formally, a Markov blanket means that, if the blanket states are known, knowing the external states would provide no additional information about the internal states, and vice versa. The most intuitive example of this would be a membrane, but a Markov blanket need not be an actual physical boundary, as it is concerned with a statistical conditional independence between states. Friston furthermore proposes to parse up the blanket states into *sensory* states, which are not influenced by the internal states, and *active* states, which are not influenced by the external states. The system of interest is then identified as the internal states plus the blanket states, termed the *particular* states. In this scheme one can identify the states over which the blanketed system has 'control', i.e. that are conditionally dependent only on the particular states, namely the internal states and the active states, which are then called the *autonomous* states (see fig. 3.3; also compare this with fig 3.1).

With this four-part partition in place, the system is intended as a model of a self-organizing system (the particular states) and its environment (the external states) in a non-equilibrium steady state with each other. In other words, it is a model of a system that keeps itself in a state of homeostasis with its environment. When one solves the deterministic flows of the separate parts of the system under this partition, what results is the autonomous states performing a gradient descent on surprisal of the particular states. This thus means that the internal and active states 'act' so as to keep the particular states within less surprising states given the NESS-density of the particular states. Equivalently, it means that the internal and active states can be seen as maximizing the evidence for the prior 'belief' that the particular states are in a non-equilibrium steady state. Since the NESS-density of the particular states defines what the particular subsystem is and how it is able to persist, it is also sometimes said that such a self-organizing subsystem can been seen as engaging in 'self-evidencing' (Hohwy 2016)—as it is maximizing the 'evidence' for its own existence. More informally—and anthropomorphically—we could then say that, given that the blanketed system embodies the 'belief' that it is the kind of system that defines what it is, it will think and act so as to maximize the evidence for the belief that it exists. This is where the cognitive interpretation of these dynamics starts to come in.

Up to this point the self-organizing homeostatic system and its environment are modeled as in a state of an eternally persisting harmonic balance, where the blanketed system is able to perfectly persist for all time. Making this a more realistic model of natural homeostatic systems is where the variational free energy is introduced. To begin, surprisal is often a computationally intractable quantity, and the surprisal of the particular states also depends on the external states (through the sensory states), which means a blanketed system would realistically not be able to 'evaluate' this quantity. Variational free energy is known to be a calculably tractable upper-bound on surprisal (i.e., is it provably always greater than or equal to surprisal). The variational free energy of the particular states furthermore only depends on the particular states, which means it *can* be 'evaluated' by the blanketed system. There are a variety of ways to derive and express this variational free energy quantity. One of them involves it parametrizing part of the difference between two probability distributions (the 'Kullback-Leibler divergence'). When expressing the variational free energy of the particular states, this quantity captures part of the difference between the probability distribution of the internal states and a true posterior probability distribution of the external states given the blanket states. When this quantity is minimized (as a proxy for surprisal), it thus equivalently looks as if the manner in which the internal and external states covary is such that the internal states come to approximate a true posterior Bayesian belief of the external states given the blanket states. In other words, the internal states can be seen as performing approximate Bayesian inference with respect to the external states conditioned on the blanket states.

Thus, substituting surprisal for variational free energy, the blanketed system comes to perform a tractable task that approximates self-evidencing as well as Bayesian inference with respect to the external states. And when the system minimizes the variational free energy of its particular states, it is effectively capable of maintaining a non-equilibrium steady state with its environment. From here on the FEP basically comes to tell the same story as predictive processing, albeit in a more abstract and statistical manner. That is, the blanketed system is able to adaptively maintain its structure by minimizing the degree to which it is surprised by causes propagating from the external states to the sensory states. This can be done by either changing the internal states, or by letting the active states change external or sensory states. Here the dynamics of the internal states and active states are thus, respectively, taken to correspond to the processes of perception and action, or rather: *perceptual inference* and *active inference*. As these two processes unfold, the internal states come to embody an approximation to the true posterior probability distribution of the external states given the blanket states, which can be understood as approximate Bayesian inference with respect to the environment. Friston thus spells out the self-organizing dynamics of this kind of system as an *inferential* or *epistemic* dynamics, where the self-maintaining system effectively performs approximate Bayesian inference, if only in an implicit sense. This Bayesian interpretation is taken directly from the manner in which free energy minimization is applied in statistics and machine learning.

Note that this FEP model of an adaptive system tells us very little about how the system came about and what it is physically, chemically, or biologically *actually* doing in order to realize what it does. The FEP simply starts from the assumption that these systems exist, and then looks at the statistics and dynamics of what such a system must be doing at an abstract level. That is, it simply starts by putting a Markov-blanketed system and its environment in a non-equilibrium steady state. It thus also assumes that there is an environment that affords the adaptive system's continued existence. Under the assumption that natural adaptive systems can indeed be modeled as Markov-blanketed systems, the FEP then purports to tells us that the fact of their existence and persistence means that their dynamics can mathematically be viewed as an inferential process.

In the FEP's original formulation in neuroscience, as underpinning PP, the nervous system is taken to correspond to the internal states, the sensory epithelia to the sensory states, and the musculature and other regulatory mechanisms to the active states, while the rest of the body and the environment are all part of the external states. One can equip these states with additional mathematical structures so as to make the perceptual and active processes more sophisticated and make such FEP models a more accurate representation of how actual nervous systems operate. Plenty of theoretical and empirical work is undertaken in this regard within neuroscience and cognitive science.[25] Friston and others have however also started to venture into the realm of theoretical biology and quite explicitly apply the FEP more generally to all living systems. Here, organisms generally are thought to embody the dynamics proposed by the FEP—or rather, they *must* do so as an existential imperative—and a nervous system is just one evolutionary addition to the dynamics of certain living systems with which the inferential processes can be performed in a more sophisticated manner. The organism as a whole is thereby taken to correspond to the particular states, and its internal physiology can be viewed as dynamically embodying a process of 'making inferences' about the environment. The FEP as such is taken to provide a model of how living systems are able to keep themselves in a state of homeostasis. Multicellular organisms are thereby viewed as consisting of a nested structure of Markov-blanketed systems—with single cells, organs, organ systems, and living systems as whole all having their respective Markov blankets.[26]

Friston and others have also compared the FEP directly to the principle of least action in physics. As in Lagrangian formulations of physics where, by specifying the boundary conditions, one can determine that a system takes the path of least (or rather, stationary) action, the FEP proposes that, given that a Markov-blanketed system maintains its structure for a non-trivial amount of time, one can determine that such a system takes the path of least variational free energy of its particular states. One could link the variational free energy minimizing dynamics to actual thermodynamic quantities using Landauer's principle.[27] From this it can purportedly be stated that a system implementing this FEP scheme is following the path of least action by information-theoretically following the 'path of least variational free energy' (Linson et al. 2018, 15). This would then, quite literally, mean that the FEP provides a principle of least action of Markov-blanketed systems. This is also part of the reason Friston insists on calling the FEP a principle. The 'equations of motion' that one can derive from this principle are then supposed to provide a physics of living and cognitive systems, or what is sometimes called 'Bayesian mechanics'

[25] See e.g. Parr et al. (2022) for an accessible overview of this work.

[26] Perhaps, intuitively, a single-celled organism best illustrates the idea of a Markov-blanketed system.

[27] Landauer's principle states that in order to erase one bit of information, an amount of energy of at least $E = k_B T \ln 2$ is required.

(Friston 2019; Kim 2021), in a manner directly analogous to how one can derive Lagrangian or Hamiltonian mechanics from the principle of least action.

Critics have pointed out that the FEP should not be seen as a theory of biological self-organization or of neurocognitive mechanisms, but rather as a mathematical framework (e.g. Andrews 2021), or a mathematical result that applies only to (weakly mixing) random dynamical systems in a non-equilibrium steady state with a Markov blanket. This qualification may not straightforwardly apply to (all) living systems, and some critics have in particular pointed out problems with conceiving of living systems as Markov-blanketed systems (e.g. Aguilera et al. 2021; Bruineberg et al. 2021; Raja et al. 2021). If these critiques are correct, it would seriously limit the scope of the FEP.[28] Proponents of the FEP on the other hand see Markov blankets as a necessary feature of any system that can meaningfully be said to 'exist', in the sense that they can measurably be distinguished from their environment. Seth (2021, 185) for examples states that "the best way to think of the FEP is as a piece of mathematical philosophy rather than a specific theory that can be evaluated by hypothesis testing", where this piece of 'mathematical philosophy' proposes to tell us what the necessary preconditions for 'existence' are in our physical world—what Colombo & Wright (2018) call the 'transcendental argument' of the FEP.[29] The FEP then draws conclusions about biological self-organization and neurocognitive mechanisms from extremely general statistical considerations regarding the viability of organisms' survival in unpredictable environments (Buckley et al. 2017, 74). As a rather abstract framework the FEP does not tell you much about specific cases unless you ask whether measurable systems conform to the principle (Friston 2018). When applied in specific cases the FEP might thus provide one with a potentially insightful general conceptual framework.

Part of the appeal of the FEP lies in the fact that it conceives of the dynamics of biological and cognitive phenomena in such a way that it seems very much compatible with our understanding of physics, and thus allows us to understand better how biological and cognitive phenomena fit into the world of physics (Seth 2021, 185). Phrases have been uttered like the FEP providing a 'physics of the mind' or a 'physics of sentience' (Ramstead et al. 2018). Since the FEP's 'Bayesian mechanics' proposes an *inferential* dynamics, the more accurate phrasing would probably be a 'physics of inference'. But insofar the FEP

[28] One could furthermore take issue with the assumption at play in the FEP that there are universal general principles at play underlying the (apparent) heterogeneous workings of living and cognitive systems (Parr et al 2022, 5). Critics of this kind of 'first-principles' approach could argue that underlying this heterogeneity may well be a disunified jumble of different kinds of processes.

[29] Andy Clark (in an interview with Closer to Truth) has also made the comparison between the FEP and string theory, in the sense that it has the ambitions of providing a 'grand unified theory' for the cognitive and life sciences, but that it is not entirely clear (yet) if and how one could empirically test this idea, and that it may turn out to primarily be an elegant mathematical framework.

provides a physics of anything, it is a physics of Markov-blanketed systems remaining (approximately) in a non-equilibrium steady state, the mathematics of which can be restated in information-theoretic terms as approximate Bayesian inference. If this is a valid account of living systems, this provides an account of the physics of certain kinds of complex adaptive systems that behave 'as if' they perform inferences. That in itself does not address 'sentience'; it merely shows how seemingly inferential behavior can be seen to emerge out of purely physical behavior.[30]

If the claims of the FEP are legitimate, and the ideas apply to the self-organizing dynamics of living systems generally, we may have a foundational physical mechanism on our hands that fleshes out the dynamics of living systems in *epistemic* terms, providing a physical and biological basis for inferential processes. Life could then be described as fundamentally engaged in 'epistemic foraging'. As Levin & Dennett (2020) put it, this allows us to make sense of life in general in terms of 'cognitive agents'. As they argue, the point of this is not to 'anthropomorphize' the blind dynamics of purely physical systems, but rather to naturalize cognition. Under the FEP we can describe living systems as agents with 'goals' and 'desires' and performing actions in service of those goals and desires, and we are arguably mathematically licensed to recognize purposeful, inferential behavior in the natural world and to adopt an 'intentional stance'. Ramstead et al. (2018) have proposed an ambitious FEP-based ontological framework for the life sciences in general, where the FEP allows one to formulate a multi-scale ontology from micro-biology to ecosystems and sociocultural phenomena, all based on (nested structures of) Markov blankets. This idea has been captured which such proclamations as 'Markov blankets all the way down' (Ramstead et al. 2018) and 'cognition all the way down' (Levin & Dennett 2020). The central idea here is that in the whole domain of life systems have a fundamental imperative to minimize free energy in order to persist, both at the cellular level and the collective behavioral level.

One could also note that the Markov blanket partition is completely symmetric, and in the formalism involved we could just as well swap around the external states and internal states. Due to this symmetry, one can similarly in the exact same fashion interpret the external states as performing approximate Bayesian inference with respect to the internal states. This symmetric perspective has led to Friston and others conceiving of the macroscopic process of evolution as a Bayesian optimization process—that is, as the environment 'inferring' and 'learning' what organisms best fit its environmental niches in the

[30] See e.g. Seth (2021) about how PP (and the FEP) could inform our thinking about consciousness, as possibly coming about in sufficiently complex cognitive systems that have sufficiently rich predictive models of themselves, a richer modeling of percepts and possible actions, and models with more temporal depth.

same way organisms themselves infer what actions and beliefs best fit their continued existence (e.g. Hesp et al. 2019, 217-220).

All of this is still rather speculative, but if all the above is legitimate, this might provide a new basis for theoretical unification within the life sciences, as well as an enormous opportunity for theoretical unification between cognitive science, biology, and physics. For naturalistically inclined philosophers the FEP might thus be a promising element for a naturalized metaphysics. It may also play a role in a naturalized epistemology. In a highly abstract, first-principled way, the FEP might provide an account of how epistemic behavior is physically possible, and how it functions physically. Friston has claimed that the FEP can be seen as a model of 'epistemology'. This should probably be understood as the claim that, given the constraints of physics, the FEP tells you what epistemic behavior can and must be like. One of the interesting issues that the FEP may help us address is how apparent goal-directed, purposeful, representational, and inferential behavior emerges out of things 'just obeying the laws of physics'. The idea that self-organizing systems in general can be described as agents performing Bayesian inference at least tells us that there is something in the physical world that we can understand as an informational dynamic between systems and their environments, which may provide us with a physical basis for epistemic behavior. Epistemology as the theory of knowledge concerns itself with the relation between agents' beliefs and the world, and what it is for this relation to be one of 'knowledge'. Insofar as the FEP might enlighten us about the form that this relation takes, it might provide us with some preliminary ingredients to build a naturalistic epistemology with by starting at the level physics.

In a recent paper Beni & Pietarinen (2021) have argued that the FEP may even allow us to naturalize 'the scientific method'—C. S. Peirce's conception of the scientific method in particular. In their paper they map elements from Peirce's formulation of the scientific method—in terms of abduction, deduction, and induction—to concepts in the FEP. Since these ideas seem to map to each other pretty well, they argue that the FEP, as a scientific 'theory' of inferential processes, may allow us to naturalize the scientific method. However, the FEP is only supposed to provide a physical and biological basis for making sense of inferential processes existing in the natural world. That does not tell us much yet about how science naturally comes about. The FEP only provides a rather minimal basis for a naturalistic epistemology. To truly make the FEP do epistemological work, it would have to be applied to cases where more complex cognitive systems can utilize this basic dynamical inferential process to form cumulative cultural constructions, normative reasoning practices, and many more of the other ingredients essential to science. The FEP cannot directly 'naturalize' those ingredients, other than providing a plausible basis for them in statistical physics and biology. The general lesson to take from the FEP is that processes we are mathematically warranted to describe as 'inferences' can demonstrably be identified with fairly straightforward physical processes, under some modeling assumptions. This tells us that 'inference' can be seen to be part of and emerge out of purely physical processes. But if one wants to go from this FEP account of an inferential dynamics to a naturalistic account of human epistemic agents doing science, one requires a number of extra steps. The FEP mechanism does not directly explain anything about the possibility of science in a physical world, it only provides some of the potential groundwork. From these 'foundations', one would have to ascend to the sociocultural realm, where science resides.

Here I have provided only a crude sketch of some of the central claims of the FEP. In essence, the FEP states that Markov-blanketed systems that persist over time, and hence approximate a non-equilibrium steady state, must minimize a variational free energy function of their particular states. The mathematics of this process can be restated in information-theoretic terms as approximate Bayesian inference with respect to external states influencing the sensory states. If living systems can indeed be modeled as Markov-blanketed systems, this may provide a physics of living systems that behave 'as if' they perform inferences, thereby suggesting a continuity between life and cognition. It thereby provides a somewhat deflationary account of cognitive behavior—merely in virtue of a system not 'falling apart' the system can be described as performing cognitive tasks. The FEP—if correct in its ambitious claims—may also provide us with a physical and biological foundation for a naturalistic theory of epistemology; but this foundation is rather minimal, as it only provides a physical and biological basis for inductive and self-correcting processes. Regarding what further role the FEP's 'physics of inference' might play in a naturalistic theory of epistemology, one would need to look to applications to the kinds of complex systems that can utilize the basic process proposed by the FEP to form such things as cumulative cultural constructions and normative reasoning practices—as I will discuss in the next chapter.

IV. THE SOCIOCULTURAL DYNAMICS OF SCIENCE

4.1 Cultural Evolution

Twentieth-century philosophers of science like Karl Popper, Thomas Kuhn, David Hull, and others, often made analogies between the workings of science and biological evolution, and used evolution as a metaphor in their philosophical account of science. The general thought there is that, in scientific communities, scientific ideas are selected for their experimental and explanatory success, and ideas that perform poorly on those scores are weeded out. The scientific community's body of accepted theories and hypotheses then 'adapts' to the natural world through a selection mechanism, which then 'evolves' to better fit that world over time.[31] This is a compelling metaphor to some extent, but there is a sense in which the attainment of knowledge and science can more explicitly be understood as an evolutionary process. These ideas fall under the banner of 'evolutionary epistemology'. Approaching epistemological questions through an evolutionary lens can be done in broadly two (complementary) ways (Bradie & Harms 2020): (i) by studying the evolution of cognitive mechanisms in animals, and humans specifically, and their reliability for epistemic purposes; and (ii) by studying the evolution of ideas, norms, and cultures in general, and how theories of cultural evolution can shine a light on epistemological questions. The topics discussed in the previous chapter may perhaps go some way in accounting for the first aspect. Here I will focus on the second aspect, concerning cultural evolution.

Regardless of what one may think at this point about the merits of the ambitious universal claims of the FEP, we can still recognize that PP in neuroscience has a more established standing, and thus we can recognize that PP may accurately account for (human) cognitive behavior. In this and the next section I will among other things briefly address some connections that have been drawn by researchers between PP (or the FEP) and the behavioral dynamics of animals, and especially to the sociocultural dynamics of humans.

As mentioned above, Beni & Pietarinen (2021) propose that the FEP may provide a way to naturalize 'the scientific method'. I would argue that such an employment of the FEP is too quick. As an aside on the terminology, it is good the keep the distinction in mind between 'the scientific method' understood as the normatively proper strategy for handling ideas by subjecting them to testing and criticism, and the *socially organized way* by which that strategy is carried out (Godfrey-Smith 2003, 224).[32] Beni & Pietarinen (2021) concern themselves with the first sense. If what is meant by naturalizing the scientific method using the FEP that this first sense—the normatively proper strategy for acquiring knowledge about the world—can be identified with a natural process in the world, then the FEP may indeed, in a rather abstract sense, allow us to naturalize 'proper' epistemic behavior. This however does not tell us how this strategy

[31] Note that in the case of Kuhn the evolutionary analogy was meant so as to be able to make sense of scientific progress, where progress should not be understood in terms of a teleological progression towards truth but rather in terms of an adaptation towards being able to better solve problems dealt by the natural world—just as organisms do not evolve towards an ideal organism; they evolve to better fit an ecological niche (see e.g. Bird (2018), section 2).

[32] Traditionally, this first sense concerned questions in the philosophy of science, whereas the second concerned questions in the history and sociology of science; in recent decades these concerns have merged considerably.

becomes implemented to gain explicit, intersubjectively ratified and justified, knowledge of the world. This latter aspect is, arguably, much more crucial for understanding what 'science' is. The real 'scientific method' is to submit ideas to social scrutiny, and through institutional selection pressures scientific communities produce reliable knowledge claims in the form of hypotheses, models, and theories. The FEP by itself does not tell us how that could happen naturalistically. Thus, if we think of science, as well as 'the' method it employs, more as a social mechanism, the question becomes how such a social mechanism can have evolved in the cultural species that we are. If one wants to use the FEP to naturalize the scientific method in that broader sense, one would have to turn to FEP applications to sociocultural dynamics, as well as existing work on e.g. cultural evolution—work that may perhaps be subsumed under it.

Assuming that we have a plausible physical story to tell concerning how (apparent) epistemic behavior may arise from simple physical mechanisms, the question remains what more complex and explicit forms this is able to take in the case of humans. A number of researchers in recent years have speculated about applying PP and the FEP to the behavioral dynamics of social animals, and especially to the sociocultural dynamics of humans. Ideas about the evolution of culture rooted in biological processes have been gaining traction in recent decades, with such notions as niche construction and gene-culture coevolution, as well as a number of theories of 'cultural evolution'. The idea behind applying the FEP to these ideas is not so much to provide an alternative framework, but rather to provide a general theoretical framework that can undergird these notions.

Hesp et al. (2019, 220-223) argue that FEP applications to animal behavior cast niche construction behavior as the process whereby organisms 'outsource' the computation of the expected uncertainty reduction under an action to the statistical structure of the physical environment. Niche construction can thus be viewed as the process whereby organisms make their ecological niche conform to their expectations. In fact, niche construction is viewed as a direct corollary of active inference, as the process by which actions are undertaken to attune the statistical structure of the environment in order to bring about expected sensory outcomes. In a collective behavioral setting, related organisms engage in a similar free energy minimizing dynamics with their shared environment. The individual 'active inference' processes acting on the (social) environment then result in (re)constructing that environment so as to collectively minimize free energy. In this interaction between related organisms and their shared environment shared expectations emerge from their collective free energy minimization.

Veissière et al. (2021, 15) argue that in the context of the behavioral dynamics of groups of organisms living together in a niche the FEP becomes a 'principle of most affordance'. Affordances are opportunities for (adaptive) actions that the environment offers to a particular organism; and for an affordance to be perceived by an organism, the organism must be able to pay attention to the right properties its environment.[33] Under the FEP, organisms select actions that have the least expected free energy, and thereby bring about preferred expected sensory outcomes, which can be associated to actions that have the most affordance. The path of least variational free energy for a living system then becomes, in the behavioral dynamics of groups of similar living systems that share an environmental niche, the 'path of most affordance'—that is, with respect to aspects of the (social) environment that are (perceivably) transformed through actions.[34] As a visual illustration of the idea, one can think of a 'desire path' (Veissière et al. 2021, 15-16) that is gradually carved out in a patch of grass, and thereby affords a (literal) path that provides opportunity for efficient movement. The idea here is that niche construction behavior in general can be thought of as 'carving out paths' in the (social) environment that optimizes affordances for the organisms in that environment.

In the case of humans, niche construction behavior takes the form of a quite radical (re)construction of the (social) environment. The types of affordances at play in the human cultural context include those that depend on the ability to cope with implicit or explicit expectations, norms, conventions, and cooperative social practices in order to better interpret other agents in a symbolically and linguistically mediated social world (Ramstead et al. 2016, 3), as well as elaborately construed cues in the physical environment. Ramstead et al. (2016) refer to such affordances as 'cultural affordances', which are characteristic of cultural species. The shared expectations that emerge from collective cultural affordance optimization are argued to induce 'regimes of shared attention' that then guide and constrain social practices, which in turn shape those expectations (Ramstead et al. 2016; Hesp et al. 2019, 217). Under this perspective, social norms can be cast as shared 'solutions' that are arrived at and learned through the collective free energy minimization of people within a culture (Colombo 2014). Social norms as shared expectations result in a degree of 'synchronization' between the members of a certain (sub)culture, which allows those members to make more accurate inferences about the expectations (and therefore likely behavior) of other members (Hesp et al. 2019, 217). As such, FEP-based sociocultural accounts propose that the kind of cultural behavior we humans engage in can be seen as a more elaborate scaffolded extension of a general biological behavioral phenomenon.

[33] The notion of an 'affordance' originates with James J. Gibson's work on ecological psychology and is a concept that is often drawn on in situated approaches to cognitive science.

[34] Unaltered aspects of the environment may of course still present affordances that are not part of this niche-construction process.

In his book on evolutionary epistemology, Harms (2004) similarly argues that human culture, as a biological phenomenon, can be thought of as essentially a collection of behavioral dispositions following from phenotypic variability. Animal species other than humans have these to varying degrees.[35] In humans however, such behavioral dispositions due to phenotypic variability have been amplified in unique ways, where they can be manifested in ways such that they can accumulate in ever more complex ways over time within and over generations. In humans many behavioral dispositions follow from imitation. Imitation, a form of high-fidelity transmission, makes it possible that behavioral dispositions are roughly shared in a population, where they can accumulate, cluster, and evolve.[36] This is the manner Harms (2004) and others argue that cultural evolution can occur. Harms argues that we can think of the evolution of cultures in behavioral terms by focusing on the coordinated state change in the behavioral dynamics in communities. He defines cultural evolution as "*the dynamics of the distribution of acquired behavioral traits among human beings, with the emphasis on the role that imitation plays in governing this dynamics*" (Harms 2004, 64) (his italics). One of the mathematical tools researchers use to model the dynamics of the proliferation of acquired behavioral traits in populations is evolutionary game theory, a tool Harms deems uniquely suited for the study of cultural evolution.

In his work on the evolution of social conventions, Skyrms (2015) similarly takes cultural evolution to be analogous to biological evolution. In the case of cultural evolution, spontaneous trials of new behaviors and recombination of complex thoughts and strategies give rise to variation (Skyrms 2015, xiv); and successful strategies and ideas are communicated and imitated more often than unsuccessful ones, which gives rise to differential replication. The population dynamics of this process can be modeled using evolutionary game theory, which can show, in an often simplified and idealized manner, how certain strategies and frequencies of them can evolve if they are evolutionarily stable. Evolutionary game theory models the dynamics of strategy change as it is influenced by the frequency of competing strategies in a population. How the relative frequency of strategies changes over time is a function of the 'fitness' associated with a strategy, or rather, the utility of that strategy to an agent in an interaction with a random other agent. The rate at which the frequency of a strategy will increase or decrease is proportional to the ratio of the strategy's

[35] As a simple illustration of this Harms (2004, 61) for example mentions cold-blooded animals that use their behavior to regulate their body temperature (cf. Damasio (2018) on (human) culture as a behavioral means to achieve homeostasis).

[36] The notion of 'ideas' as the elements of cultures that evolve led to the concept of 'memes' as discrete informational entities analogous to genes (e.g. Dawkins (1976), Dennett (1995, 2017). This 'meme' concept has faced substantial criticism, as it is unclear what such memes are supposed to be and what their ontological status is (e.g. Sterelny 2006, Sperber 2001, Harms 2004).

fitness to the mean. In the case of an evolutionarily stable configuration the population dynamics is attracted to certain points or regions of the configuration space where it will eventually tend to settle. Using evolutionary game theory, Skyrms shows how, in admittedly idealized situations, certain social norms could have evolved, or must have evolved, without getting into more complex realistic scenarios.

Evolutionary game theory can thus be used in order to simulate what sorts of strategies and norms might evolve out of the iterative interactions between agents in a population. Here too it has been argued that the FEP may undergird game-theoretic accounts of behavior (Friston 2010). In game-theoretic models agents always optimize perceived utility, which can, in FEP terms, be understood in terms of minimizing expected free energy under an action. The FEP might then provide a mechanism by which game-theoretic scenarios (or other agent-based models) can be realized in the natural world.

Regarding the relevance of evolutionary game-theoretic models of cultural evolution to epistemology, Skyrms (2015, 81-105) for example discusses a simple scenario in which meaningful language, in the form of a signaling convention (whereby a signal corresponds to a certain event) can spontaneously evolve. Skyrms takes the case of vervet monkeys, who have four signals to warn for different predators, as an inspiration for simulating the population dynamics of such a 'signaling game'. Specifying all the sixteen possible strategies, Skyrms shows that in virtually all initial population distributions one of the two actual signaling strategies eventually takes over the population. Thus, in such cases Skyrms shows that via several pathways 'meaning' can spontaneously evolve. Although the evolution of human language is obviously much more complicated, Skyrms takes this example to show that at the very least it is possible for meaningful language to evolve.

Regarding evolutionary epistemology, Harms (2004) shows how in evolutionary game-theoretic models generally, information about the environment in a measurable sense (i.e. 'mutual information') always increases in a population over time (in a constant environment). That is to say, in the evolutionary population dynamics the state of a population always tracks the state of the environment in such a way that the state of the population becomes more informative about the environment. One can directly compare this to Fisher's theorem (Okasha 2005), which states that the mean fitness of a population will always increase by natural selection (again, in a constant environment).[37] As populations are 'climbing a fitness landscape', the state of

[37] Note that this caveat of a constant environment means that fitness, or information, *can* go down if the environment changes. In that case, however, populations will always start to increase their fitness with respect to this new environmental state, and thus also gain information about the new environment.

the population covaries with the state of the environment such that the information-theoretic measure of mutual information increases in the state of the population, and the population thus ‘acquires’ information about the environment. Harms takes this to imply that we can view the evolutionary dynamics of life and culture in terms of a process of information-transfer from the environment to populations, in which there is always an information gain about the environment in populations. In the case of cultural evolution then, the evolutionary dynamics transfers information about the world to our ideas (“across the Kantian barrier”, as Harms (2004, 183) puts it). As such, Harms argues he can quantifiably show something that we ordinarily assume (in everyday life or in epistemology)—namely that our ideas reliably track things in the external world and are thus informative about the world (Harms 2004, 182). Harms concludes from this demonstration, as well as applications of it to a number of case studies (such as the evolution of social norms), that “natural relationships exist which largely have the form we believe the objective rules of reason and behavior have” (Harms 2004, 243). This is reminiscent of the sort of claims made by proponents of the FEP, and it can be directly compared to Beni & Pietarinen’s (2021) claims regarding the FEP’s relevance to a naturalized epistemology. But here too Harms’ claims about how information about the world can increase in populations and cultures provides only a very minimal naturalistic basis for how ‘knowledge’ and science could evolve in a culture. If we want to naturalize epistemology, we would require more than such a minimal basis.

It should also be noted that the types of examples that Harms and Skyrms provide are extremely simple, idealized, and tractable situations, and one can object that these simple cases cannot really do much epistemological work unless they are extended to much more realistic cases that are much messier and more complicated—as well as to more specifically epistemological cases (Okasha 2005).[38] The work to be done is then to see if these simpler cases can be extended to more complex and realistic cases that, admittedly, are many orders of magnitude more difficult to model—as these would incorporate a hugely dimensional space of possible strategies, and more complicated payoffs and costs associated with those strategies.[39] Furthermore, one could object that these processes of cultural evolution as discussed by Harms and Skyrms in ‘normal’ circumstances give shape to things like the manifest image and all

[38] Okasha (2005) objects that Harms’ story about mere information-theoretic covariation between population states and the environment does not yet tell us much about the epistemological issue of how *beliefs* can be accurate *representations* of the external world. For my purposes, I could refer back to the FEP in that case.

[39] In the field of complexity science some researchers work precisely on such complex cases with a variety of agent-based modeling techniques. Such more realistic cases may thus be on their way (or may well already exist in the complexity science literature).

manner of cultural phenomena, but not necessarily science. The types of cases discussed so far may tell us something about the seed for scientific knowledge gathering, but they apply to culture generally. The question regarding epistemology and science becomes how cultural evolution could also have given rise to scientific practices and the scientific image. Over the history of the human species culture has evolved to all manner of immense complex interconnected forms like global political and economic institutions, advanced technology, and science. What might we argue 'demarcates' science as a special kind of cultural evolutionary innovation? That will be the topic of the following section.

4.2 Reason & Criticism

Philosophers of yore characterized humans as the 'rational animal'. It is hard to deny that there is something particularly special and different about the human animal, seeing as we now unprecedentedly dominate most of the planet. But what might we say makes the human species so unique in the animal kingdom? There is probably an indeterminate list of characteristics that one could identify, but we can at least pinpoint three crucial aspects, namely (i) hyper-sociality, (ii) dexterity, and (iii) a relatively large and flexible, modifiable brain, which respectively results in humans having (accumulative) culture, the capacity for complex tool-use, and a more sophisticated form of general intelligence. These three aspects and their consequences are most likely tightly interwoven in the evolutionary history of the human species, and they likely amplified each other. We need not get into the detailed evolutionary history of the human species here; instead we may ask how these human characteristics may have given rise to human 'rationality'.

Rationality is often understood as the deployment of reason, where reason would be the capacity for making valid and sound deductive or inductive inferences. If this is what is meant by rationality and reason, then humans are generally not particularly rational animals. A more interesting proposal for understanding 'reason' has recently been put forth by Mercier & Sperber (2017) in the form of their argumentative theory of reason.

Let us start by recognizing again that humans are an inherently cultural species. The human brain is extremely flexible and largely *needs* to be 'shaped' by cultural upbringing, resulting in the extraordinary long human infancy.[40] The

[40] This is not to say that the human mind is a *tabula rasa*—a blank slate—that is completely formed by sociocultural conditioning. There is still much that is 'pre-wired' in the human mind, albeit much less than in other animal species. If anything, the way in which humans can acquire or learn behavioral dispositions, i.e. be ' nurtured', is severely biased by our biological and psychological make-up, i.e. our 'nature'.

evolutionary strategy that the human species hit upon was, instead of having everything be ‘wired up’ from the get-go, to let this animal come into the world ‘half-baked’ and let the (social) environment wire the animal up.[41] This has turned out to be an incredibly successful strategy that has made the human species enormously adaptive to many different environments, most of which it now dominates. This strategy is of course not without risk, because it entirely relies on human infants receiving the proper (cultural) upbringing, which in some tragic cases they do not. But this risky strategy has served the human species enormously well, with modern human science and technology being recent additional triumphs that were gained from it.

But how did humans achieve these recent triumphs? Humans, so the story goes, have been able to achieve this because they have the capacity to reason. But that capacity for reasoning hangs tightly together with the manner in which humans are an inherently cultural and hyper-social species. Mercier & Sperber (2017) propose that reason ought not to be seen as some kind of general-purpose cognitive superpower—a cognitive add-on. This supposed superpower, as is generally recognized in psychology, is systematically flawed and biased; and the idea of a flawed superpower makes very little evolutionary sense—a function cannot have evolved for something it systematically malfunctions for. Mercier & Sperber instead propose that reason, in the sense of the particular human cognitive capacity for *reasoning*, is a cognitive skill that particularly evolved for the purpose of navigating the complex social environment that humans found themselves in. Reasoning evolved for the purpose of providing *reasons* for one’s behavior and ideas to others, and conversely to evaluate the reasons given by others. This ‘argumentative theory of reason’ thus states that, evolutionarily speaking, reason in humans serves the purpose of providing and evaluating arguments, which was a crucial capacity to have in the hyper-social niche that humans built for themselves. Reasoning, i.e. exchanging and evaluating arguments, allowed human social groups to negotiate their collective behavior in highly adaptive ways. Explaining human reasoning in this way, Mercier & Sperber claim, also explains many of the cognitive biases that are inherent in our reasoning capacities.[42] Thus, insofar humans are ‘rational animals’, they are so because reason evolved as a social skill so as to be deliberating animals.

[41] This colorful expression is borrowed from the neuroscientist David Eagleman (see his appearance on episode 122 of the Mindscape podcast).

[42] That is, the biases in human reasoning are generally such that it “overwhelmingly finds justifications and arguments that support the reasoner’s point of view” and “makes little effort to assess the quality of the justifications and arguments it produces”, which makes a lot of sense “for a cognitive mechanism aimed at justifying oneself and convincing others” (Mercier & Sperber 2017, 9-10).

The perspective put forth by Mercier & Sperber fits well into a more situated and interactionist understanding of (human) cognition.[43] And as Veissière et al. (2020, 4) argue, this interactionist model of human cognition lends itself well to a culturally informed FEP model. That is, it fits well into the active inference framework of an inferential dynamics with the environment, where for humans the hyper-social niche they constructed for themselves became an intrinsic part of that environment. The inferential dynamics of the individual agents with this social niche led to a more explicit form of inference that humans needed to communicate to other members in their social environment in a self-reflective and metacognitive way. A communicative effort is of course an action, and thus—under the FEP—falls under the rubric of an 'active inference'. Providing an argument is then a special instance of active inference—one that becomes more explicit through communicative and linguistic means. Humans then had to be able to evaluate 'arguments' provided by others, and in so doing contribute inferences to other structures in their predictive models of the environment (as well as to themselves as structures that make inferences).[44] This may furthermore also have driven the need for developing an elaborate communication system, i.e. language, which furthermore allowed for more cognitive outsourcing and scaffolding into the (social) environment.[45] It has furthermore been argued that most silent reflective thinking going on inside human heads may more aptly be thought of as an exaptation of, or perhaps even a rehearsal for, argumentation with, and justifications to, others (Veissière et al. 2020, 16).[46]

Clearly, we do need to distinguish between Mercier & Sperber's *descriptive* account of human reason—which provides an account of how human reasoning capacities have evolved, what they have evolved for, and thus are adapted to—and a *normative* or *prescriptive* account of 'ideal' reason—which is concerned with the identification of valid or otherwise correct and appropriate arguments by studying them with the aid of e.g. formal systems of logic, mathematics, and probability theory. Mercier & Sperber argue that argumentation has evolved in humans in order to better exploit their social environment by persuading others into thinking alike with convincing arguments. But what is considered

[43] It also aligns well with the 'social brain hypothesis' which proposes that "the fact that primates have unusually large brains for body size compared to all other vertebrates" is because "primates evolved large brains to manage their unusually complex social systems" (Dunbar 2009, 562).

[44] It has been argued that this metacognitive capacity has made humans hypersensitive to attributing intentionality and agency to things in their environment, which may be part of the reason all human cultures developed some form of religion (e.g. Atran (2002) and Dennett (2006)).

[45] That is, the noises and gestures produced during speech acts, as sound waves and optic phenomena, of course become environmental cues and affordances for other agents (as well as for the agent that produced them).

[46] An exaptation is a 'repurposed' trait—one that evolved for a certain function but later came to serve another.

convincing by the human mind is not the same thing as being correct. A common definition of a fallacy is an argument that seems to be valid or appropriate but is not so. The fact that an erroneous argument seems correct to a person is of course a consequence of human psychological biases and a variety of social circumstantial factors. What we can say however is that the descriptive account may tell us how it is possible that a certain subset of humans (e.g. early scholars, philosophers, academics, and later, scientists) came up with *normative* accounts of 'proper' reasoning. So even though the way most humans in fact reason is quite different from, and quite imperfect according to, normative ideals of reason, we can understand how this real form of reasoning was capable of constructing higher ideals of reason, for example with the aid of additional cultural tools such as the formal linguistic systems of logic and mathematics.[47] These external tools then allowed humans to improve their reasoning and 'extend' and outsource their cognitive processes of reasoning into their (social) environment.

Science is of course often heralded as *the* pinnacle of human reason. With science, human reason has been able to 'uncover the secrets of nature' and use that to the great benefit (or detriment) of the human species. Science as a cultural innovation can be seen as having constructed a novel social environmental niche in which sociocultural behavior and reasoning practices are highly constrained to evolve in only very limited ways—ways that turned out to be extremely successful for a variety of purposes. Such types of constraints can be seen to have become embodied in the social organization and process that arose with the advent of science. The way that science as a social mechanism organized itself was, among other things, to institutionalize norms of criticism. In the case of science, communities have hit upon a strategy that changes the rules of the game of ordinary cultural evolution. By instantiating strong norms of criticism and thus weeding out errors and weak arguments, the social niche in scientific communities introduces different kinds of environmental selection pressures for sociocultural constructs. Science as a social process allows its 'culture' to evolve in a very constrained manner in accordance with certain epistemic norms towards containing ever more accurate, clear, predictive and unified ideas. And if we were to agree with Harms (2004) and Beni & Pietarinen (2021), scientific norms can be seen as an actual reflection of processes in the natural world.[48] This sociocultural innovation of science happened fairly

[47] Note that I view mathematics here primarily as an especially precise and rigorous language that humans have developed for the purpose of the unambiguous communication of ideas and arguments and the modeling of natural and imaginary phenomena.

[48] We may indeed take the normatively proper way of generating knowledge about the world to be more or less dictated by the structure of the world itself. That is, the natural world itself constrains how information (e.g. ideas, models, theories) can come to accurately represent the natural world—there are

recently, even though elements of that innovation have existed in other cultures in the past, and exist in other 'subcultures' other than science—and different fields of science may embody that innovation in slightly different ways. The precise historical details as to why this cultural innovation occurred around the sixteenth century in Europe are not of great importance here.[49]

To illustrate how the argumentative theory of reason may also enlighten our thinking about science, Mercier & Sperber draw on a well-known example in the history and philosophy of science, namely that of Newton's engagement in alchemy. Mercier & Sperber use this to emphasize how Newton's activity in these domains can be considered to *not* have been very 'rational' and scientific. Precisely because these activities were solitary and shrouded in secrecy, they provided no pressure to produce strong arguments or anticipate counterarguments—in contrast to his work in mathematics and physics, which was part of a discussion in a public and critical community. As Mercier & Sperber (2017, 327) put it: "When reasoning about gravity, Newton had to convince a community of well-informed and skeptical peers. He was forced to develop better arguments. When reasoning about alchemy, there were no such checks. The same brilliant mind reasoning on its own went nowhere". This is put forth to emphasize that science is not about 'solitary geniuses' propelling the process forward, but rather more about critical communities in which reasoning faces proper checks and balances. It is the very embeddedness of individual scientists in the social structure of science, and the access to the finely crafted tools designed for cognitive scaffolding that makes science the pinnacle of human reason.[50] There is a lot of cultural scaffolding at play in science, and the cognitive extensions that are constructed in scientific communities allow for the attainment of higher ideals of reason. By 'cognitive extensions' I mean anything from systems of logic and mathematics and the existing bodies of accepted scientific knowledge, to the error-filtering mechanisms inherent in the social structure of science. Science could thus be seen as a kind of social mechanism that constrains human reasoning, funneling and amplifying it, and through the institutionalization of criticism creates strong demands for the cogency, empirical adequacy, consistency, and rigor of ideas, in addition to whatever other epistemic virtues one might identify. This of course brings us into the

only limited ways in which representations can evolve subject to these constraints. The FEP and evolutionary selection processes may indeed allow us to naturalize this aspect.

[49] A non-exhaustive list of elements of this innovation may include a certain combination of empiricist and rationalist philosophical ideas, the rejection of ancient or religious authorities, the dictum to 'take no one's word for it' (*nullius in verba*) and 'see for oneself' through experimentation and cogent argumentation, the mathematization of the study of nature, norms of accepting and anticipating criticism, and (later) the institutionalization of scientific activities.

[50] As Daniel Dennett (2017) likes to say: "You can't do much carpentry with your bare hands, and you can't do much thinking with your bare brain."

territory of classic topics in the philosophy of science, namely questions regarding what 'science' is, what makes something 'scientific', and what the epistemic virtues of science are or should be.

As should have become clear by now, I do not think science should be regarded as simply a more sophisticated form of normal every-day reasoning, as some philosophies of science do. Following the insights of Kuhn and others, we can recognize that science is a particular kind of social mechanism, one that is particularly well-equipped for 'solving problems' and producing reliable knowledge. The specific mechanism of 'paradigms', 'normal science', and 'paradigm shifts' that Kuhn proposed may not have been wholly accurate, but later thinkers have taken up and attempted to improve on this account of science as a social mechanism. Lakatos and Laudan for example stressed that we should think of scientific fields as 'research programs' or 'research traditions', whose research is to be progressive if it is to be considered properly scientific, meaning that the scope, precision, and internal consistency of the theory and auxiliary hypotheses has to increase over time. Synthesizing the insights from Popper, Kuhn, Lakatos, and others, Thagard (1978) for example proposed that we demarcate science by looking at a combination of the internal consistency and empirical adequacy of the theory, the critical attitude of the community surrounding the theory towards inadequacies of the theory, and the historical context regarding other (competing) research programs.[51] A discipline or theory can then be said to be 'scientific' if the community engaged in it scores well on these three features, on the third score meaning that it does better than alternative theories at the time.

The defining characteristic of science may thus much more be the critical attitude of scientific communities—the manner in which norms of criticism are embodied in the social niche of science. This might be one of the most central features of science, as one might expect that logically and mathematically valid 'arguments' and empirically adequate theories would naturally emerge from such a community. As Ladyman (2013, 56) puts it, science is about the "emergent properties of the scientific community and the interactions among its members, as well as between them and their products." Science can thus be understood as adding constraints to the way reasoning 'naturally' operates in human societies, and through those constraints cultural evolutionary processes have been able to tap into a whole new domain of adaptive cultural constructions: those of empirically adequate, predictive, instrumental, and explanatory theories of the world.

With the variational free energy minimizing inferential dynamics and predictive cognitive processes discussed previously, amped up in complexity

[51] Such approaches to the demarcation problem have more recently been discussed further in Pigliucci & Boudry (2013).

and made more explicit due to the complicated demands for social cognition, and the accumulative character of human cultural evolution, we can start to make sense of human intelligence giving rise to science. The last step we still need to take is to understand how the cultural affordances constructed by scientific communities in the form of linguistic constructs in both natural and artificial languages (i.e. mathematics)—that is, scientific theories—are in any way representative of their real world. I do not intend to rehearse the old debate about scientific realism here; rather, we can ask the more deflationary question of how the naturalistic epistemic agents doing science, that we model inside the scientific image, might 'correctly' track and represent the world of the scientific image. This will be the topic of the next and final section.

4.3 Tracking Real Patterns

Metaphysics as a discipline is in the business of developing basic criteria for determining what sorts of 'things' should be considered 'real'. As I discussed earlier in chapter 2, Ladyman & Ross argue that a naturalistic way of doing metaphysics should be in the service of showing "how the separately developed and justified pieces of science (at a given time) can be fitted together to compose a unified world-view" (L&R, 45). Ladyman & Ross themselves argue that, despite their insistence on the primacy of physics, our current scientific image is highly suggestive of an emergentist ontology, or what they call a 'scale relativity of ontology'. In their view, a proper naturalistic theory of ontology should incorporate all sorts of entities and properties that are not necessarily 'physical', because many powerful explanations and successful predictions—referring to these entities and properties—have come forth from sciences that are not physics (L&R, 41). They spell out this scale relativity of ontology in more detail by formulating their ontological theory of 'real patterns'. This theory of real patterns is then taken to be an example of how one might actually *do* metaphysics in a naturalistic manner.

For my purposes here, the theory of real patterns is of relevance as a proposed ontological framework of the scientific image. Tentatively going along with this real-patterns conception of our scientific worldview, I wish to argue that we can make sense of communities of scientists, as so far described, as capable of latching onto the actual ontology of their world, and thereby accurately represent their world. In their elaborate discussion of the mind-independence of real patterns, Ladyman & Ross for example argue that usefulness to possible observers and the capacity to grant inductive and explanatory success can be, and often is, an indicator for a *real* pattern, and not merely of instrumental or pragmatic value. This is the main point I wish to stress here in this last section.

Crudely put, the idea behind 'real patterns' is to recognize the ontological status of 'emergent' phenomena—to see them as *real* and not mere useful coarse-grained descriptions for the purpose of epistemic book-keeping. The metaphysical picture that results is one opposed to an eliminativist or strong reductionist one. In a sense, a real patterns theory concerns a proposal for a definition or measure for when some 'higher-level' process can truly be said to be 'emergent', and to thereby be ontologically committed to emergent phenomena at all scales. The notion of real patterns that Ladyman & Ross build on originates in Daniel Dennett's 1991 paper 'Real Patterns'. In his paper Dennett tries to argue for a realist attitude towards higher-level emergent phenomena by appealing to the information-theoretic notion of a pattern. In that case there are non-instrumental facts of the matter about which patterns are present in a dataset. Information-theoretically, a pattern is just the presence of regularities, and thus redundancies, in a dataset that can be exploited to compress the dataset, and Dennett (1991, 34) argued that "a pattern exists in some data—is real—if there is a description of the data that is more efficient than the bit map, whether or not anyone can concoct it." Ladyman & Ross however criticized Dennett's paper for being unclear about whether the 'real patterns' should be regarded as *real* or as useful fictions. In their attempt at providing a more rigorous and comprehensive account of what real patterns are they then take on an explicitly realist attitude towards real patterns.

Information-theoretically, one could thus say that a pattern in some dataset is anything other than pure randomness, and that the dataset thus contains regularities, and hence redundancies that can be exploited to compress the dataset. Such a compression is then analogous to a description or 'theory' of a higher-level phenomenon. More generally, we could say a pattern is any kind of (statistical) regularity that can be captured by some projectible, generalizable rule (or 'theory'). Ladyman & Ross basically argue that the *real* patterns may be defined in terms of the *most efficient* way of coarse-graining some system—one that provides the most predictive leverage and explanatory power—about which there would be an objective fact of the matter.[52] The ontology of the scientific image could then be defined in terms of those unique ways of coarse-graining the world—truly 'carving nature at its joints'.[53]

[52] Although making reference to maximum efficiency sounds like a anthropocentric or subjective criterion, Ladyman & Ross argue that there is a physical fact of the matter about the computational effort saved with employing a real pattern description of a system. Computation is a physical process, and as such there are objective physical facts about what these descriptions are (Ladyman et al. 2013).

[53] As such, Ladyman & Ross take the natural world to have an objective modal structure—that is, to have objective relations of necessity, possibility, potentiality, and probability between events. There being an objective truth to the (statistical) laws capturing this modal structure at different levels renders their position 'anti-Humean' with respect to the metaphysics of natural laws—a discussion I will not get into here.

Obvious examples that are supposed to be captured by a real patterns theory are systems that have some dynamical stability at a certain timescale, like e.g. atoms or molecules.[54] Clearly, carving up the world along the lines of atoms and molecules provides an enormous predictive and explanatory leverage, compared to describing the world only along the lines of for example quarks, bosons and electrons.[55] What would then make atoms and molecules *real* patterns is the extent to which models of atoms and molecules most parsimoniously capture what goes on at the spatiotemporal scales where collections of fundamental particles form identifiable dynamically stable configurations. Thus, by throwing out all the microphysical details, but keeping a compressed description of the pattern at a higher scale, one is able to retain an enormous amount of predictive power against only a relatively small loss in accuracy (Ladyman et al. 2013, 64). Making the quantities involved in such an endeavor precise, and then optimizing them, is what a theory of real patterns attempts to do regarding the identification of *real* patterns. Criteria for real patterns are then supposed to capture all the *bone fide* scientific ontological categories, where e.g. 'horses' are considered real patterns, but 'unicorns' are not.[56] Of course, where it starts to become increasingly contentious is what the real patterns above the scales of atoms and molecules are—what kind of patterns are real for example in economics or ecology? The hope is that complexity science might be of help in better discerning what sorts of patterns may be considered real in those domains (Ladyman et al. 2013).[57] And if for example the FEP is indeed an adequate account of the dynamics of living systems, it may in fact identify a class of real patterns in the domain of complex adaptive systems at a variety of scales.

There are different ways of construing this real patterns theory (Ladyman 2017, 154). One could on the one hand posit a two-tier ontology with the world of fundamental physics plus all the real patterns that emerge from it at higher levels (e.g. Wallace 2003); on the other hand one could also go 'all-the-way' and think of existence across the board in terms of real patterns, where emergent

[54] That is, at the timescale of trillions of years, atoms and molecules are not stable.

[55] It is worth pointing out that quantum mechanically speaking, atoms and molecules are *dynamical processes*, not 'things' or 'objects'—and thus *patterns* in the physical structure of the world. The same goes for elementary particles, which are themselves excitations in quantum fields, and thus also more appropriately thought of as structural patterns.

[56] I take this example from an interview with James Ladyman on the Mindscape podcast (episode 33).

[57] Ladyman & Ross also argue that "the kind of processes fecund with real patterns are those that exchange thermodynamic entropy for information at high rates" (L&R, 240). The real patterns production rate in a region can be thought of as a non-linear function of the thermodynamic depth of that region at a specific scale. That would then be why the special sciences have such a rich ontology to uncover here on Earth. Real patterns in the special sciences must be stabilized by something against entropic dissolution, that is, a disposition to resist background perturbations. This would be its extent to which a pattern supports projectibility by physically possible observers. Therefore, "as a pattern's stability goes asymptotically to zero, it ceases to be real", or, differently put: "to be (in this context, to persist) is to resist entropy" (L&R, 250).

phenomena are on equal ontological footing with those of putative fundamental physics. This second construal is the one Ladyman & Ross settle for, and thus they take their formulation to truly imply a scale-relativity of ontology. For my purposes here it does not really matter which construal to side with. Of course, we know that our current theories of physics cannot be the end the of the story.[58] Whatever new more fundamental theories of physics end up being like, they *must* recover what current theories predict accurately, and thus our current theories must be effective theories that are consequences of them in the appropriate regime. Our current theories of physics may thus be understood as reflecting real patterns.[59]

A real patterns theory of ontology as an ontology of the scientific image is thus an attempt to unify the theories developed and entities identified by different scientific disciplines into one metaphysical framework. Ladyman & Ross's initial formulation of this theory was in terms of information-theory, where a real pattern could be defined in terms of the most efficient compression of some dataset—in line with Dennett's original formulation. This particular formulation has received criticism on a number of both technical and conceptual grounds (e.g. Beni 2019; Suñe & Martínez 2019), and Ladyman (2017, 153) later also suggested that real patterns may be defined in terms of non-redundant statistical structures (e.g. Ladyman & Ross 2013), or in terms of the dynamics of phase-spaces, where real patterns allow a reduction of the degrees of freedom. What these different formulations have in common is, as I put it above, to provide a measure for a uniquely efficient coarse-grained description of some system, and to then recognize *that* as the ontologically *real* pattern in the world. By attempting to provide a unique formal measure, a real patterns theory is proposed as capturing the objective invariant structure of the world at different scales. Compare this with the famous remark by Einstein in a 1933 lecture that "it can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience"—which is sometimes paraphrased as 'everything should be made as simple as possible, but no simpler'. In a sense, this is a restatement of Occam's principle of parsimony. A real patterns theory can be understood as turning this principle into a multi-scale theory of ontology.

[58] Our most fundamental theories of physics, quantum field theory and general relativity, are known to conflict or 'break down' in certain regimes, thus we know they cannot be accurate theories of fundamental physics. A testable unified theory of 'quantum gravity' has thus far eluded physicists.

[59] For all we know, maybe physics does not 'bottom out', and it is 'real patterns all the way down'—a possibility seriously considered by Ladyman. Also note that non-fundamental theories of physics, like Newtonian mechanics, statistical mechanics, or thermodynamics, are also thought to capture real patterns under this view.

To again make an explicit connection with the previous sections, there is a sense in which one might also find an application of the FEP to a real patterns theory. As argued by Beni (2019), the FEP or PP may be of use for developing a formal measure or definition of what real patterns are. Building on work by Churchland (2012), Beni argues that modern theories in neuroscience and biology can be used to explicate the representational relationship between scientific theories and the structure of the world. Beni's novel contribution is to incorporate PP and the FEP into such an account. Beni's own proposal is a position he terms 'cognitive structural realism', where the real structures in the world are *defined with respect to* cognitive processes existing in that world. Beni's primary motivation for formulating his brand of structural realism is to address and defuse skeptical considerations regarding scientific representation.[60] Such issues are not of much relevance for my purposes here. Working from *within* the scientific image, I already assume the 'truth' of physics and other scientific theories, including—tentatively—the FEP.[61] I have deliberately avoided tricky metaphysical debates about meaning and representation here. Under the assumption that the world is structured in the manner described in physics, epistemic agents and their products are part of that physical structure, and as such they can be conceived of as devising representational tools that possess a structure-preserving mapping to the invariant structure of their world in some manner. Internal to the physical world, we can understand perfectly well how such a mapping might work.

Beni's radical solution to overcoming skepticism about scientific representation is more or less to short-circuit it by simply saying that what exists is what can be defined with respect to the kinds of quantities invoked by PP or the FEP—which would mean cognitive processes are by definition in the business of representing the structure of the world. This is not so much what I intend to do here, but it is worthwhile to consider some of his arguments for this move, as his explanation for this move is one that I *can* appeal to in making sense of scientific research programs latching onto real patterns (as more generally thought of by Ladyman and others initially).

[60] See e.g. Clark (2017) for arguments in a similar vein against skeptical implications following from PP and the FEP (as put forth by Jakob Hohwy (2016; 2017)). Clark's argument is, roughly, that Markov blankets are not static 'veils' through which internal states infer the in principle unreachable external world; rather, they are dynamical processes that continually reconstitute themselves and are thus coupled to the external world and the internal states. This dynamical coupling between internal and external states would then undermine an evil-demon-like skeptical scenario.

[61] This is not to say that (epistemic) agents within the scientific image would not face e.g. underdetermination problems—they certainly do. But the epistemic values of parsimony undergirded by the real patterns theory would allow us to say that the selection mechanisms of theory-choice in fact are nudged in the 'right' direction.

Beni criticizes Ladyman & Ross's initial information-theoretic definition of a real pattern as being an abstract and purely formal measure, and states that Ladyman & Ross's account "fails to tell a convincing story about how to naturalize information or ground physical patterns in the physical world" (2019, 80). This would in part be because the proposed information-theoretic measure of a real pattern would not be uniquely definable for some dataset, and as such would depend on the practical interests of a researcher (Beni 2019, 76). Beni proposes we may use PP to properly 'naturalize' measures of 'real patterns' or 'structures', as discussed in the structural realism literature—that is, to ground the (information-theoretic) measures one might use to define them in the physical world. Since under PP the brain is argued to work like a Bayesian inference engine that attempts to infer the statistical and causal structure of the world around it, Beni argues that such cognitive mechanisms can be used to specify the 'structures' of the world (Beni 2019, 6). The reason for this move is that, under PP and the FEP, cognitive systems have a natural propensity for latching onto the world in virtue of the continual minimization of prediction errors (or 'free energy'). As such, a statistical inferential link is forged between cognitive systems and the causal structure of the world, which provides a viable basis for ascribing reliable representational powers to the brain (Beni 2019, 121). Viability constraints and evolutionary selection pressures should furthermore increase our confidence that the representational relationship must be a reliable one.[62] The informational structure of neurological representations, as well as scientific representations, may then be explicated in terms of embodied informational structures that are entwined with causal structures in the physical world (Beni 2019, 9).

Without getting into the details regarding Beni's proposal, I only wish to take his proposal as an indication that researchers have been thinking along similar lines as I am doing here, and that we can properly talk about cognitive inferential processes *in* the physical world that represent statistical regularities in that physical world—and as such there is a sensible way in which we can start talking about epistemic agents 'accurately' representing their world. Of course, much of the representational structure proposed in PP and the FEP resides implicitly and unconsciously in the wisdom of organisms' bodies.[63] The question I raised in the previous two sections is then how we can make sense of scientific communities explicitly representing their world accurately. Human beliefs and products do indeed take on this more explicit representational character in virtue of

[62] Since variational free energy is an upper bound on surprisal, there is no guarantee that the representational relationship will be a truly faithful one—the bound is not necessarily zero. Nevertheless, there are viability constraints and evolutionary selection pressures that keep this bound as small as possible.

[63] Or for that matter: in the organizational structure of social groups.

cumulative cultural evolution, argumentation and language. But that does not tell us yet whether they represent the world 'as it really is'. For that they would have to accurately capture the real patterns in the world.

Human ideas can thus reasonably be said to represent things in the world, but in normal circumstances that happens through a pragmatic trade-off between accuracy, usefulness, and evolutionarily and developmentally acquired 'priors' and psychological inclinations that introduces biases. In scientific communities this pragmatic trade-off is significantly skewed towards accuracy and a critical revision of some of the priors, and thus arguably has a better chance of getting nearer to the real patterns in the world. If those real patterns are defined in terms of uniquely and maximally efficient descriptions, through criteria of parsimony, efficacy, projectibility, and communicability, we would essentially be defining them in terms of the survivability of criticism. It then makes sense that scientific communities are, under ideal circumstances, able to track, approximate, and sometimes accurately represent those real patterns.[64] Ideas that latch on to real patterns are the only ones that can survive in the long run in cultural niches of continual criticism.[65]

In this way we can think of the scientific culture of criticism as an environmental selection effect for ideas to become aligned with the real patterns of the world. Scientific cultures come to embody empirically identified epistemic norms and virtues, in virtue of which we can also define the real patterns of the world. Occam's principle of parsimony is in fact also argued to be a fundamental heuristic in the inferential dynamics of the FEP—that is, natural systems conform to this principle in their self-maintaining (and 'inferential') behavior. Variational free energy can be expressed as a combination of, roughly, an hypothesis' degree of overfitting and the lack of generalizability (Mann et al. 2021, 11). In minimizing variational free energy a balance is struck between the two in a parsimonious manner (settling for an 'hypothesis' that makes the least amount of assumptions so as to account for most of the data), and as such Occam's principle is implemented. By extension to applications to culture, science thus also embodies this principle, albeit in a more streamlined fashion. In a recursive manner we could then thus define the ontology of the world in a real patterns theory with reference to the manner in which cognitive systems operate (i.e. in line with Occam's principle), which would mean that epistemic behavior in a sense can be defined with respect to the ontic structure of the

[64] I say 'under ideal circumstances', as it is certainly the case that certain social, political, or economic factors can make the critical social environment in a scientific community less than ideal—as sociologists of science well know.

[65] This is, of course, because if a scientific idea at hand is *not* in fact the most parsimonious account of some piece of empirical data, critics may eventually fault the idea for this and propose other ideas that purportedly are. Over time, scientific ideas are then expected to come closer to the defining criteria for real patterns.

world, and vice versa. That is, there would be a certain symmetry between the inferential dynamics of epistemic agents existing in a world and the ontology of their world. There becomes a close connection between epistemology and ontology. Boldly put: if to be is to be a real pattern, then what is exists is what epistemic agents track, and what epistemic agents track is what exists. In line with Beni's 'cognitive structural realist' position, we can think of real patterns precisely in terms of those types of statistical regularities that cognitive systems (according to the FEP) have a natural tendency to latch onto. 'Tracking' or 'latching onto' is however not the same as accurately representing and explicitly knowing the real patterns of the world—most of the inferential dynamics is implicit, pragmatic, and approximate. I have argued here that we can conceive of scientific communities as capable of actually explicitly representing what exists approximately or accurately through particular sociocultural mechanisms.

Science, when working properly, can be seen as a type of social organization that captures objective invariant features of the world, features that could be defined with a theory of real patterns. As I argued in sections 3.1 and 3.2, we can understand this as the construction of cultural affordances through a collective free energy minimizing process—which in science takes on a very particular form. We could then understand the relation between the world and successful scientific theories to be one of a structure preserving mapping between the world and those theories that are themselves part of that world. As embedded, embodied, and dynamical features of the physical world itself, we may thus understand epistemic agents engaged in science and their products as complex dynamical aspects of the world that track, approximate, and accurately represent the real patterns of the world.

As I pointed out earlier, my point here is not to formulate a form of scientific realism, or to rehash that ongoing debate in the philosophy of science. The view expressed here is not intended to be scientific realism of the 'traditional' kind, but rather one of a more deflationary kind. Scientific realism, or variants thereof, can be taken as a 'scientific hypothesis' to account for the empirical success of scientific theories. 'Scientific realism' may then be viewed as part of an account of naturalistic epistemic agents whom we model within our scientific image. Consider the following quote from Peter Godfrey-Smith (2003, 229):

> Scientific belief is not the product of us alone or of the world alone; it is the product of an interaction between our psychological capacities, our social organization, and the structure of the world. The world does not "stamp" beliefs upon us, in science or elsewhere. Still, science is responsive to the structure of the world, via the channel of observation.

From the perspective *within* the scientific image that I have been taking in this thesis, we can understand this as naturalistic epistemic agents being capable of tracking the objective invariant structure of their world.[66] I hope I have been able to sketch out what this 'interaction between our psychological capacities, our social organization, and the structure of the world' might look like according to a collection of contemporary scientific and philosophical ideas. The last step sketched out in this section has been to propose that the ontology of our scientific image can be regimented in terms of real patterns that can naturally be tracked by cognitive processes (as proposed by PP and the FEP). These real patterns can furthermore be explicitly represented, approximately or accurately, by sociocultural constructs engineered by human scientists.

V. CONCLUSION: BACK AGAIN?

In the introduction I briefly discussed the case of the Eddington's two tables, and I noted that a 'naturalistic' approach to resolving the tension between the two different perspectives on the table was to start with the 'physical table', and then from there work back to something resembling the 'manifest table'. The idea there is that with a rich enough arsenal of scientific theories—ranging from physics to psychology—we would ideally be able to tell a convincing reconciliatory story about the two tables that would make the manifest table a lot less surprising given the physical table. I have taken on a similar kind of perspective in this thesis, and I have tried to apply it more generally to the philosophy of science. My aim with this thesis was to articulate how one might go from a scientific worldview—where the world is taken to 'ultimately' be physical—to epistemic agents that can construct that very scientific worldview. In that regard, I hope I have been to tell a somewhat convincing reconciliatory story.

The reconciliation sought for in this thesis was between our more 'manifest' understanding of what science is and how it works, and our 'scientific' understanding of science and scientists as ultimately physical processes. I have tried to argue that the scientific image provides ample material to work with to account for epistemic agents doing science that then construct the scientific image. As such, we can view the scientific image as being perfectly self-consistent. Trying to demonstrate that self-consistency is to a large extent what I take a naturalistic philosophy of science to consist of. As I argued in chapter 2,

[66] From a 'meta-perspective' we can understand this as a statement about the scientific image being responsive to and constrained by the real world (which would land us into an actual discussion on scientific realism).

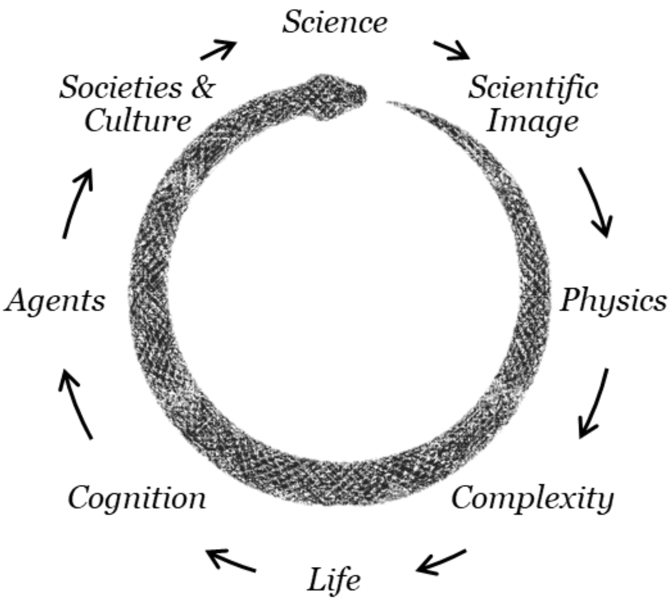

Fig. 5.1: An illustration of the recursive kind of naturalistic philosophy of science argued for in this thesis.

we can think of naturalism in the philosophy of science in terms of a kind of 'virtuous circularity', where philosophers work *from within* the scientific image to make sense of science. In trying to come up with a self-consistent view concerning how science fits into the scientific image, I believe we are essentially formulating a naturalistic philosophy of science (see e.g. fig. 5.1). The main argument in the background of this thesis is perhaps much more that there is plenty material to draw on in our current scientific image that would allow one to spell out such a convincing reconciliatory story.

What I have done in this thesis is to provide merely an example of how one might give shape to such a story. In my account I have made use of a number of ideas that I have been exposed to over the years and find of interest. Clearly, one may disagree with a number of the elements I have drawn upon in my story on a number of different grounds. The main point I hope to have made, however, is that something like this *can* be done—the scientific image is rich enough and novel emerging scientific insights provide plenty of the required material such that one can spell out such an account. Thus, even if one disagrees to a large extent with many or most of the elements that I have discussed in this thesis, my point is perhaps much more that there is plenty material to draw from in which one can tell a self-consistent story of the scientific image. Perhaps one may thus go a different way, by using an entirely different set of ideas, and hopefully still arrive at a similar conclusion. The notion that *that* aim is feasible is much more the general argument of this thesis—that with a self-consistent story of the scientific image at hand, one should be able to formulate a naturalistic theory of epistemology and naturalistic philosophy of science.

I may also add that with such a naturalistic philosophy of science we also very much have grounds to take the scientific image of the world seriously. Sometimes sociologists of science, expounding more relativist and social constructivist views, take themselves to actually take a more proper naturalistic approach to science, and taking the idea to its proper conclusion (Godfrey-Smith 2003, 221). My account in this thesis can in part be seen as a counterargument to that, where I think a proper naturalistic philosophy of science can certainly be sympathetic to some kind of 'realism'. In a naturalistic approach to science we can certainly make sense of science as representing the 'real world' in some sense.

The naturalistic philosophy of science as I formulated it—going from physics to epistemic agents—does not necessarily suggest any way of *doing* actual science. It much more suggests that the current scientific worldview that appears to emerge out of the body of scientific theories about the world is perfectly consistent with there being scientists doing science and 'discovering' how the physical world operates—or 'constructing' theories and models that accurately account for the workings of the physical world. The scientific image of the world—as I have presented it—provides us with ample and perfectly natural dynamical processes that can account for the possibility of epistemic agents. What is presented here is thus more akin to a kind of 'existence proof'—the main point being that the scientific image can incorporate epistemic agents constructing the scientific image. This means we can, and should, continue elaborating the scientific image within the current scientific paradigm, without the need for a radical revision—as some philosophers and scientists at times suggest. The physical universe of our scientific image is causally closed, and its internal causal structure is enough to account for ourselves as epistemic agents studying it.

APPENDIX: A SKETCH OF THE FREE ENERGY PRINCIPLE

In the following I provide a crude summary of some of the central claims of the FEP—skipping over many technical details—primarily based on Friston (2019), Parr et al. (2019), and Raja et al. (2021).[67]

The FEP starts with the assumption that the system of interest can be modeled as a *weakly mixing random dynamical system*, whose behavior can be described with a stochastic differential equation (or Langevin equation):

$$\dot{x} = f(x) + \omega \tag{1}$$

Here, the rate of change of the system x—where x defines all the relevant state variables—is determined by a deterministic flow $f(x)$ plus random perturbations ω.

Given a Langevin system, the time evolution of the probability density function $p(x)$ of this system—i.e. a continuous distribution describing the relative probability of finding the system in a certain state x at a particular time—can be described with the Fokker-Planck equation:

$$\dot{p}(x) = \nabla \cdot \big(\nabla \Gamma p(x) - f(x)p(x)\big) \tag{2}$$

Here, Γ is the amplitude of the random fluctuations ω.

Simply setting this equation equal to zero means the probability density remains static over time, in which case the system is also said to be in a 'non-equilibrium steady state'. In this static form the deterministic flow $f(x)$ can be solved as follows:

$$\dot{p}(x) = 0 \iff f(x) = (\Gamma - Q)\nabla \ln p(x) \tag{3}$$

Written here in the form of a 'Helmholtz decomposition', this flow can be decomposed into a curl-free (irrotational) flow component f_Γ and a divergence-free (solenoidal) flow component f_Q:

$$\begin{gathered} f_\Gamma = \Gamma \cdot \nabla \ln p(x) = -\Gamma \cdot \nabla \Im(x) \\ f_Q = -Q \cdot \nabla \ln p(x) = Q \cdot \nabla \Im(x) \\ \Im(x) = -\ln p(x) \end{gathered} \tag{4}$$

The irrotational flow component f_Γ counters the dispersion of the density, that would otherwise be caused by ω, by performing a gradient ascent on the logarithm of the probability density (see fig A.1). The solenoidal flow component f_Q (as a function of an antisymmetric matrix $Q = -Q^T$) circles around the

[67] There are many equivalent ways of formulating and arriving at some of the central concepts and ideas in the FEP, especially in specific applications, but here I focus on its most basic form, or what Parr et al. (2022) call the 'high road'.

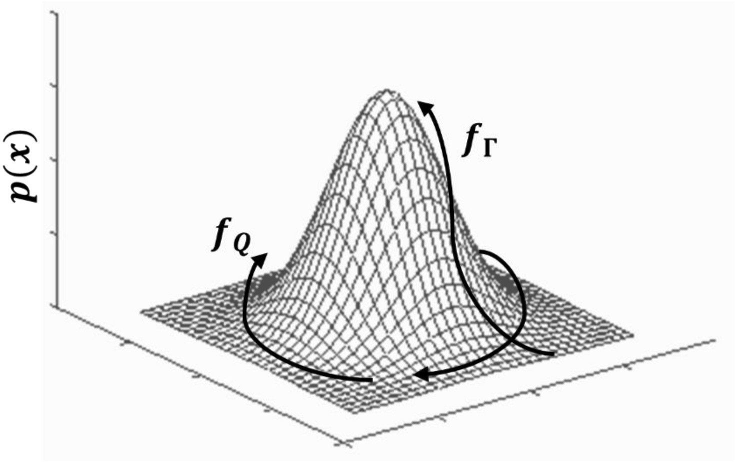

Fig. A.1: Gradient ascent on a probability density function.

contours of the logarithm of the density; since it is a circular current, the density remains the same everywhere. In information theory $-\ln p(x)$ is also known as the *Shannon information* or *surprisal*, termed $\Im(x)$, which is a measure of the 'unexpectedness' of some state x given $p(x)$. The f_Γ component can thus also be seen as performing a gradient *descent* on surprisal.

The next step concerns partitioning the system x such that one can identify internal states μ that are conditionally independent from external states η due to the presence of a Markov blanket b. Formally, a Markov blanket means that, if the blanket states b are known, knowing η would provide no additional information about μ, and vice versa:

$$\eta \perp \mu \mid b \Leftrightarrow p(\eta, \mu|b) = p(\eta|b)p(\mu|b) \tag{5}$$

A further step concerns parsing b up into sensory states s, which are not influenced by μ, and active states a, which are not influenced by η (see fig. A.2). The system of interest consists of μ and b which are called the particular states π. In this scheme one can identify the states over which the subsystem π has 'control', i.e. that are only conditionally dependent only on π, namely μ and a, which are called the autonomous states α. With these partitions ($x = \{\eta, s, \mu, a\}$) and their groupings ($b = \{s, a\}$, $\pi = \{\mu, b\}$, $\alpha = \{\mu, a\}$) in place one can write down the Langevin dynamics for these separate states:

$$\dot{x} = f_i(x) + \omega_i = \begin{bmatrix} \dot{\eta} \\ \dot{s} \\ \dot{\mu} \\ \dot{a} \end{bmatrix} = \begin{bmatrix} f_\eta(\eta, b) + \omega_\eta \\ f_s(\eta, b) + \omega_s \\ f_\mu(\mu, b) + \omega_\mu \\ f_a(\mu, b) + \omega_a \end{bmatrix} \tag{6}$$

Solving for these separate flows by setting the Fokker-Planck equation equal to zero—as in eqs. (3) & (4)—yields the following solution:

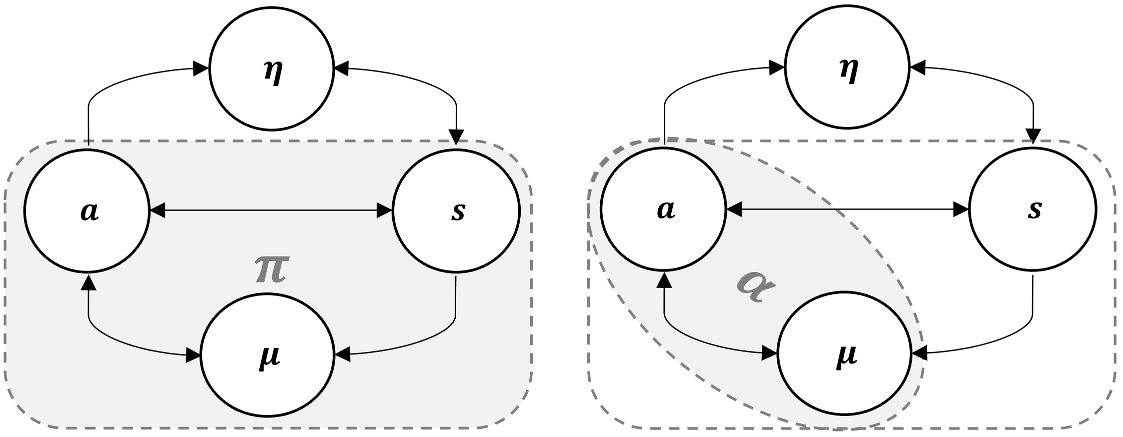

Fig. A.2: A causal Bayesian graph depiction of a Markov blanket (s, a), particular states (π), and autonomous states (α).

$$f_i(x) = \begin{bmatrix} f_\eta(\eta, b) \\ f_s(\eta, b) \\ f_\mu(\mu, b) \\ f_a(\mu, b) \end{bmatrix} = \begin{bmatrix} (Q_{\eta\eta} - \Gamma_{\eta\eta}) \cdot \nabla_\eta \Im(\eta, b) \\ (Q_{ss} - \Gamma_{ss}) \cdot \nabla_s \Im(\eta, b) + Q_{sa} \cdot \nabla_a \Im(\eta, b) \\ (Q_{\mu\mu} - \Gamma_{\mu\mu}) \cdot \nabla_\mu \Im(\mu, b) \\ (Q_{aa} - \Gamma_{aa}) \cdot \nabla_a \Im(\mu, b) + Q_{as} \cdot \nabla_s \Im(\mu, b) \end{bmatrix} \tag{7}$$

The part of the dynamics that is conditionally independent of the external states, i.e. the autonomous dynamics f_α (ignoring the solenoidal coupling between active and sensory states) looks as follows:

$$f_\alpha = (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha}) \cdot \nabla_\alpha \Im(\pi) \tag{8}$$

The flow of the autonomous states f_α thus evolves on a gradient of surprisal of the particular states . The irrotational component of this flow therefore performs a gradient descent on surprisal of the particular states:

$$f_{\alpha,\Gamma} = -\Gamma_{\alpha\alpha} \cdot \nabla_\alpha \Im(\pi) \tag{9}$$

When a system with a Markov blanket partition is in a non-equilibrium state, what the particular subsystem itself does is thus minimize the surprisal of the particular states.

As a model for what a self-organizing adaptive system must do to persist this however introduces problems, since surprisal is often a computationally intractable quantity. Natural blanketed systems would thus realistically not be able to minimize surprisal directly. The quantity called variational free energy F is known to be a tractable upper bound on surprisal ($F(\mathrm{x}) \geq \Im(\mathrm{x})$). By letting the system minimize variational free energy as a proxy for surprisal, it comes to perform a more realistic tractable task.

There are a variety of ways to demonstrate that variational free energy is an upper bound on surprisal, and ultimately it is a consequence of 'Jensen's inequality'. Here I will simply show how it also follows from some of the quantities already introduced.

Because the blanket states mediate between the external states and the internal states one can define an arbitrary mapping relation between η and μ and their dynamics conditioned on b such that the internal states μ can be described as instantiating a recognition density $q_\mu(\eta)$ of η:

$$q_\mu(\eta) \approx p(\eta|b) = p(\eta|\mu, b) = p(\eta|\pi) \tag{10}$$

This recognition density $q_\mu(\eta)$ can be understood as a density of Bayesian beliefs that approximates the true Bayesian posterior concerning the external states given the particular states. The difference between the recognition density and the true posterior can be quantified in terms of the Kullback-Leibler Divergence D_{KL}:

$$D_{KL}\left(q_\mu(\eta) \parallel p(\eta|\pi)\right) = \int q_\mu(\eta) \ln \frac{q_\mu(\eta)}{p(\eta|\pi)} d\eta \tag{11}$$

The variational free energy of the particular states $F(\pi)$ can be defined as follows:

$$F(\pi) \equiv \int q_\mu(\eta) \ln \frac{q_\mu(\eta)}{p(\eta, \pi)} d\eta \tag{12}$$

In that case the D_{KL} measure can be re-expressed as a combination of the variational free energy and surprisal:

$$D_{KL}\left(q_\mu(\eta) \parallel p(\eta|\pi)\right) = F(\pi) + \ln p(\pi) = F(\pi) - \Im(\pi) \tag{13}$$

The variational free energy can thus also be defined as the surprisal plus a bound provided by the KL divergence:

$$F(\pi) = \Im(\pi) + D_{KL}\left(q_\mu(\eta) \parallel p(\eta|\pi)\right) \tag{14}$$

Since the KL divergence D_{KL} is always non-negative, the variational free energy is an upper bound on surprisal:

$$D_{KL} \geq 0 \rightarrow F(\pi) \geq \Im(\pi) \tag{15}$$

Variational free energy is a function of the recognition density $q_\mu(\eta)$ and a generative model $p(\eta, \pi)$, which is a joint probability distribution of η and π. These quantities, it is argued, *can* be ‘evaluated’ by a particular subsystem. Letting that system minimize variational free energy as a proxy for surprisal, the autonomous dynamics of the blanketed system looks as follows:

$$f_\alpha = (Q_{\alpha\alpha} - \Gamma_{\alpha\alpha}) \cdot \nabla_\alpha F(\pi) = \begin{bmatrix} f_\mu \\ f_a \end{bmatrix} = \begin{bmatrix} (Q_{\mu\mu} - \Gamma_{\mu\mu}) \cdot \nabla_\mu F(\pi) \\ (Q_{aa} - \Gamma_{aa}) \cdot \nabla_a F(\pi) \end{bmatrix} \tag{16}$$

The irrotational component of the autonomous dynamics thus performs a gradient descent on variational free energy:

$$f_{\alpha,\Gamma} = -\Gamma_{\alpha\alpha} \cdot \nabla_{\alpha} F(\pi) = \begin{bmatrix} f_{\mu,\Gamma} \\ f_{a,\Gamma} \end{bmatrix} = \begin{bmatrix} -\Gamma_{\mu\mu} \cdot \nabla_{\mu} F(\pi) \\ -\Gamma_{aa} \cdot \nabla_{a} F(\pi) \end{bmatrix} \tag{17}$$

As variational free energy is minimized, $q_{\mu}(\eta)$ also becomes a better approximation to $p(\eta|\pi)$, which can be understood as approximate Bayesian inference with respect to the external states. This minimization, and thereby 'inferential process', happens either through the dynamical evolution of the internal states or the active states. Going along with the FEP's unified account of perception and action, and thereby cognition, eq. (17) can be seen as *the* equation of perception and action, where the irrotational flows of the internal states and active states, respectively, are taken to correspond to the processes of perception and action, or rather: *perceptual inference* and *active inference*.

REFERENCES

Aaronson, S, S. M. Carroll & L. Ouellette (2014). "Quantifying the Rise and Fall of Complexity in Closed Systems: The Coffee Automaton." Preprint. *arXiv*:1405.6903.

Aguilera, M., B. Millidge, A. Tschantz, C.L. Buckley (2021). "How particular is the physics of the free energy principle?", *Physics of Life Reviews* (article in press).

Andrews, M. (2021). "The math is not the territory: navigating the free energy principle", *Biology & Philosophy* 36:30.

Arkoudas, K. & S. Bringsjord (2014). "Philosophical Foundations" in *The Cambridge handbook of artificial intelligence*, eds. K. Frankish & W.M. Ramsey, 34-64. Cambridge: Cambridge University Press.

Ashby, W. R. (1957). *An Introduction To Cybernetics*. London: Chapman & Hall.

— (1960). *Design For a Brain: The Origin of Adaptive Behaviour*. London: Chapman & Hall.

Atran, S. (2002). *In Gods We Trust: The Evolutionary Landscape of Religion*. New York: Oxford University Press.

Beni, M.D. (2019). *Cognitive Structural Realism: A Radical Solution to the Problem of Scientific Representation*. Cham: Springer Nature Switzerland.

Beni, M. D. & A.V. Pietarinen (2021). "Aligning the free-energy principle with Peirce's logic of science and economy of research", *European Journal for Philosophy of Science* 11(94): 1-21.

Bruineberg, J., K. Dolega, J. Dewhurst, M. Baltieri (2021). "The Emperor's New Markov Blankets", *Behavioral and Brain Sciences,* 1-63.

Buckley, C.L., C. S. Kim, S. McGregor, A.K. Seth (2017). "The free energy principle for action and perception: A mathematical review", *Journal of Mathematical Psychology* 81: 55–79.

Beer, R. D. (2014). "Dynamical systems and embedded cognition" in *The Cambridge handbook of artificial intelligence*, eds. K. Frankish & W.M. Ramsey, 128-148. Cambridge: Cambridge University Press.

Bird, A. (2018). "Thomas Kuhn" in *The Stanford Encyclopedia of Philosophy (Spring 2022 Edition)*, ed. E. N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2022/entries/thomas-kuhn/>.

Bradie, M. & W. Harms (2020), "Evolutionary Epistemology" in *The Stanford Encyclopedia of Philosophy (Spring 2020 Edition)*, ed. E. N. Zalta, URL = <https://plato.stanford.edu/archives/spr2020/entries/epistemology-evolutionary/>.

Callender, C. (2017). *What Makes Time Special?* Oxford: Oxford University Press.

Carroll, S. (2016). *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself*. New York: Dutton.

Churchland, P. M. (2012). *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge: The MIT Press.

Clark, A. (2015). "Embodied Prediction" in *Open MIND*: 7(T), eds. T. Metzinger & J.M. Windt. Frankfurt a.M.: MIND group.

— (2015). "Predicting Peace: The End of the Representation Wars" in *Open MIND*: 7(R), eds. T. Metzinger & J.M. Windt. Frankfurt a.M.: MIND group.

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

— (2017). "How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds" in *Philosophy and Predictive Processing*, eds. T. Metzinger & W. Wiese, ch. 3. Frankfurt am Main: MIND Group.

Colombo, M. (2014). "Explaining social norm compliance. A plea for neural representations", *Phenomenol. Cogn. Sci.* 13, 217–238.

Colombo, M. & C. Wright (2018). "First principles in the life sciences: the free-energy principle, organicism, and mechanism", *Synthese* 198: 3463–3488.

Conant, R. C. & W. R. Ashby (1970). "Every Good Regulator Of a System Must Be a Model of That System", *Int. J. Systems Sci.* 1(2): 89-97.

Dayan, P., G. E. Hinton & R. M. Neal (1995). "The Helmholtz Machine", *Neural Computation* 7: 889-904.

Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter*. New York: W.W. Norton & Company.

Dennett, D. C. (1991). "Real Patterns." *The Journal of Philosophy* 88(1): 27-51.

— (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

— (2006). *Breaking The Spell: Religion as a Natural Phenomenon*. New York: Penguin.

— (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York: W. W. Norton & Company.

Dunbar, R. I. M. (2009). "The social brain hypothesis and its implications for social evolution", *Ann. Hum. Biol.* 36(5): 562-72.

Eddington, A. S. (1928/1968). *The Nature of the Physical World*. Ann Arbor: The University of Michigan Press.

England, J. L. (2013). "Statistical physics of self-replication", *The Journal of Chemical Physics* 139: 121923.

— (2020). *Every Life Is On Fire: How Thermodynamics Explains the Origins of Living Things*. New York: Basic Books.

Van Fraassen, B.C. (2002). *The Empirical Stance*. New Haven & London: Yale University Press.

Friston, K. (2019). *A free energy principle for a particular physics*. Preprint. *arXiv*:1906.10184.

Friston, K., M. Fortier, D. A. Friedman (2018). "Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston", *ALIUS Bulletin*, 2: 17-43.

Van Gelder, T. & R. F. Port (1995). "It's About Time: An Overview of the Dynamical Approach to Cognition" in *Mind as Motion: Explorations in the Dynamics of Cognition*, eds. R.F. Port & T. van Gelder. Cambridge: MIT Press.

Gell-Mann, M. & J. B. Hartle (1990/2018). "Quantum Mechanics in the Light of Quantum Cosmology" in *Complexity, Entropy, and the Physics of Information*, ed. W. H. Zurek, 425-458. Boca Raton: CRC Press.

Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago & London: The University of Chicago Press.

Harms, W. F. (2004). *Information and Meaning in Evolutionary Processes*. New York: Cambridge University Press.

Hartle, J. B. (2005). "The physics of now", *American Journal of Physics* 73(2): 101-109.

— (2016). "Why Our Universe Is Comprehensible." Preprint. *arXiv*:1612.01952

Hesp, C., M. Ramstead, A. Constant, P. Badcock, M. Kirchhoff & K. Friston (2019). "A Multi-scale View of the Emergent Complexity of Life: A Free-Energy Proposal" in *Evolution, Development and Complexity: Multiscale Evolutionary Models of Complex Adaptive Systems*, eds. G.Y. Georgiev, J.M. Smart, C.L.F. Martinez & M.E. Price, 195-227. Cham: Springer Nature Switzerland.

Hoffman, D. D. (2019). *The Case Against Reality: Why Evolution Hid the Truth from Our Eyes*. New York: W.W. Norton & Company.

Hohwy, J. (2013). *The Predictive Mind*. New York: Oxford University Press.

— (2016). "The self-evidencing brain." *Noûs* 50(2): 259–285.

— (2017). "How to entrain your evil demon", in *Philosophy and Predictive Processing*, eds. T. Metzinger & W. Wiese. Frankfurt am Main: MIND Group.

Hume, D. ([1739] 1978). *A Treatise of Human Nature*. Vol. 2. L. Selby-Biggs, ed. Oxford: Oxford University Press.

Ismael, J. (2015). "On Whether the Atemporal Conception of the World is Also Amodal", *Analytic Philosophy* 56(2): 142–157

— (2017). "Passage, Flow, and the Logic of Temporal Perspectives" in *Time of Nature and the Nature of Time*, eds. C. Bouton & P. Huneman. Springer: Boston.

Kauffman, S. A. (2019). *A World Beyond Physics: The Emergence and Evolution of Life*. New York: Oxford University Press.

Kim, C. S. (2021). “Bayesian mechanics of perceptual inference and motor control in the brain”, *Biological Cybernetics* 115: 87–102.

Kincaid, H. (2013). “Introduction: Pursuing a Naturalist Metaphysics” in *Scientific Metaphysics*, eds. D. Ross, J. Ladyman & H. Kincaid, 1-26. Oxford: Oxford University Press.

Ladyman, J. (2013). “Toward a Demarcation of Science from Pseudoscience” in *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*, eds. M. Pigliucci & M. Boudry, 45-60, Chicago: University of Chicago Press.

— (2017). “An Apology for Naturalized Metaphysics” in *Metaphysics and the Philosophy of Science*, eds. M.H. Slater & Z. Yudell, 141-161. New York: Oxford University Press.

Ladyman, J., J. Lambert & K. Wiesner (2013). “What is a Complex System?”, *European Journal for Philosophy of Science* 3: 33-67.

Ladyman, J. & D. Ross (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.

— (2013). “The World in the Data” in *Scientific Metaphysics*, eds. D. Ross, J. Ladyman & H. Kincaid, 108-150. Oxford: Oxford University Press.

Ladyman, J. & K. Wiesner (2020). *What Is a Complex System?*. New Haven: Yale University Press.

De Lange, F. P., M. Heilbron & P. Kok (2018). “How Do Expectations Shape Perception?”, *Trends in Cognitive Sciences* 22(9): 764-779.

Levin M., D. C. Dennett (2020). “Cognition all the way down”, *Aeon: https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas*

Linson A., A. Clark, S. Ramamoorthy, K. Friston (2018). “The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition”, *Front. Robot. AI* 5(21):1-22.

Mann, S. F., R. Pain & M. Kirchhoff (2021). “Free Energy: A User's Guide” [Preprint] URL: http://philsci-archive.pitt.edu/19961/

Mitchell, M. (2009). *Complexity: A Guided Tour*. New York: Oxford University Press.

Mercier, H. & D. Sperber (2017). *The Enigma of Reason: A New Theory of Human Understanding*. Cambridge: Harvard University Press.

Okasha, S. (2005). “Review of William F. Harms, *Information and Meaning in Evolutionary Processes*”, *Notre Dame Philosophical Reviews* 12.

Parr T., L. Da Costa & K. Friston (2019). “Markov blankets, information geometry and stochastic thermodynamics”, *Phil. Trans. R. Soc. A* 378: 1-13.

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

Raja V., D. Valluri, E. Baggs, A. Chemero & M. L. Anderson (2021). “The Markov blanket trick: On the scope of the free energy principle and active inference”, *Physics of Life Reviews* (article in press).

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Pigliucci M. & M. Boudry, eds. (2013) *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*. Chicago: University of Chicago Press.

Quine, W. V. O. (1969). “Epistemology Naturalized”, in *Ontological Relativity and Other Essays*, W.V.O. Quine, 69-90. New York: Columbia University Press.

Ramstead, M. J. D. , S. P. L. Veissière & L. J. Kirmayer (2016). “Cultural Affordances: Scaffolding Local Worlds Through Shared Intentionality and Regimes of Attention”, *Front. Psychol.* 7: 1090.

Ramstead M. J. D., P. B. Badcock & K. J. Friston (2018). “Answering Schrödinger’s question: A free-energy formulation”, *Physics of Life Reviews* 24: 1-16.

Saunders, S. (1993). “Decoherence, Relative States, and Evolutionary Adaption”, *Foundations of Physics* 23(12): 1553-1585.

— (1995). “Time, Quantum Mechanics, and Decoherence”, *Synthese* 102: 235-266.

Savitt, S. F. (2012). “Of Time and the Two Images”, *Humana.Mente Journal of Philosophical Studies* 21, 57-68.

Schuster, P. (2011). “Is there a Newton of the blade of grass? The Complex Relation Between Mathematics, Physics, and Biology”, *Complexity* 16(6): 5-9.

Schrödinger, E. (1944/2013). *What is Life?*. Cambridge: Cambridge University Press.

Sellars, W. (1962/1963). “Philosophy and the Scientific Image of Man” in *Science, Perception and Reality*, W. Sellars, 1-40. New York: The Humanities Press.

Seth, A. K. (2015). “The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies” in *Open MIND*: 35(T), eds. T. Metzinger & J.M. Windt. Frankfurt a.M.: MIND group.

— (2019). “Being a Beast Machine: The Origins of Selfhood in Control-Oriented Interoceptive Inference” in *Andy Clark and His Critics*, eds. M. Colombo, E. Irvine & M. Stapleton, 238-253. New York: Oxford University Press.

— (2021). *Being You: A New Science of Consciousness*. London: Faber & Faber.

Skyrms, B. (2015). *The Evolution of the Social Contract* (2nd ed.) Cambridge: Cambridge University Press.

Suñé, A. & M. Martínez (2019). "Real patterns and indispensability." [Preprint] URL: http://philsci-archive.pitt.edu/id/eprint/16190/

Thagard, P. R. (1978). "Why Astrology is Pseudoscience", *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1: 223-234.

Veissière S. P. L., A. Constant A., M. J. D. Ramstead, K. J. Friston & L. J. Kirmayer (2020). "Thinking through other minds: A variational approach to cognition and culture", *Behavioral and Brain Sciences* 43: 1–75.

Verhaegh, S. (2018). *Working From Within: The Nature and Development of Quine's Naturalism*. New York: Oxford University Press.

Wallace, D. (2003). "Everett and Structure." *Studies in the History and Philosophy of Modern Physics*, 34: 87-105.

Walsh, K. S., D. P. McGovern, A. Clark, R. G. O'Connell (2020). "Evaluating the neurophysiological evidence for predictive processing as a model of perception", *Annals of the New York Academy of Sciences* 1464(1): 242-268.

West, G. (2017). *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies*. New York: Penguin Press.