Synthese (2018) 195:2519–2540
https://doi.org/10.1007/s11229-016-1100-6

S.I.: PREDICTIVE BRAINS

Autopoiesis, free energy, and the life–mind continuity thesis

Michael D. Kirchhoff[1]

Received: 11 February 2016 / Accepted: 15 April 2016 / Published online: 28 April 2016
© Springer Science+Business Media Dordrecht 2016

Abstract The life–mind continuity thesis is difficult to study, especially because the relation between life and mind is not yet fully understood, and given that there is still no consensus view neither on what qualifies as life nor on what defines mind. Rather than taking up the much more difficult task of addressing the many different ways of explaining how life relates to mind, and vice versa, this paper considers two influential accounts addressing how best to understand the life–mind continuity thesis: first, the theory of autopoiesis (AT) developed in biology and in enactivist theories of mind; and second, the recently formulated free energy principle in theoretical neurobiology, with roots in thermodynamics and statistical physics. This paper advances two claims. The first is that the free energy principle (FEP) should be preferred to the theory of AT, as classically formulated. The second is that the FEP and the recently formulated framework of autopoietic enactivism can be shown to be genuinely continuous on a number of central issues, thus raising the possibility of a joint venture when it comes to answering the life–mind continuity thesis.

Keywords Life–mind continuity · Free energy principle · Autopoiesis · Dark room · Causal explanatory relations · Biopsychism

1 Introduction

This paper addresses the life–mind continuity thesis. In its strongest formulation, this thesis states that mind (cognition) is foreshadowed in life. So, the relation between life and mind must be such that both share the same basic set of organizational prop-

✉ Michael D. Kirchhoff
kirchhof@uow.edu.au

[1] Department of Philosophy, University of Wollongong, Wollongong, Australia

erties (Clark 2001; Godfrey-Smith 1996; Stewart 1996; Thompson 2007). How best to understand the nature of the continuity between life and mind is still an unresolved issue. My primary concern is not to provide an exhaustive account. Here, the life–mind continuity thesis is approached through a critical comparison of two prominent ways of illustrating the relation between life and mind. The first is the *theory of autopoiesis* (AT) developed in biology (Maturana and Varela 1980) and in enactivist philosophy of mind (Di Paolo 2009; Thompson 2007; Varela et al. 1991; Varela 1997). The second is the *free energy principle* (FEP) formulated in theoretical neurobiology with roots in thermodynamics and statistical physics (Friston 2005, 2009, 2011, 2012a, b, 2013; Friston and Stephan 2007; Friston et al. 2012).

In the philosophical literature, no systematic work has yet been done to critically compare the theory of AT and the FEP. Yet both proposals have explanatory ambition. For example, Maturana and Varela state: "the notion of AT is necessary and sufficient to characterize the organization of living systems" (1980, p. 82; italics omitted). They go further claiming that "Living systems are cognitive systems, and living as a process is a process of cognition" (1980, p. 13). Hence, if these authors are correct, the AT unifies life and mind by establishing that autopoiesis is both necessary and sufficient for life and mind. Friston is equally ambitious about the unifying and explanatory power of the FEP. He says: "If one looks at the brain as implementing [free energy minimization], nearly every aspect of its anatomy and physiology starts to make sense" (2009, p. 293). Add to this that for Friston the FEP proposed for living systems aims to give a single framework within which to unify action, perception and learning (Friston 2010).

Before going on to analyze in which ways the FEP and the AT differ, and which of the two frameworks should be preferred, let me start by making clear that the FEP and enactivist developments of AT share a number of important projects. In particular, both frameworks aim to understand the basic conditions of life; they seek an explanation of the relationship between perception and action as constituted in organismic embodiment; and can be understood as driving a shift away from orthodox internalism to a view of minds as realized in dynamics crisscrossing brain, body, and world.

Given this, the obvious, albeit potential, value of both proposals is that they aim to bring life and mind together in a single, overarching and unifying, framework, via a raft of shared explanatory interests.

1.1 Overview

This paper begins (Sect. 2) by addressing entropy—a measure of uncertainty—and its relation to two different kinds of free energy. Given that all living systems must keep an upper bound on entropy, accounting for the relation between entropy and free energy will prove of central importance to what is to come in this paper. In Sect. 3 I explore differences and similarities between the AT and the FEP. My treatment of the AT here considers its original formulation in Maturana and Varela (1980). I will show that the FEP and the AT overlap as well as diverge in their commitments about the conditions required for life. Yet, as we shall see, the FEP is ultimately better suited to

address the life–mind continuity thesis. In Sects. 4 and 5, I advance two reasons for why we should prefer the FEP to the AT, as classical formulated (Maturana and Varela 1980). Finally, in Sect. 6, I turn to explore where the FEP and recent developments of autopoietic enactivism come together as well as giving a brief discussion of their potential weaknesses and strengths.

1.2 Arguments and some implications

The FEP and the AT ground the notion of mind in the wider context of life and adaptivity. In this sense, the FEP accords with the AT. Yet the two positions are not identical. There are differences between the two frameworks in their approach to the question about the unity of life and mind. In critically comparing the FEP to the AT, this paper develops two argument for why the FEP is better equipped to explain the life–mind continuity thesis than the AT, as classically defined by Maturana and Varela (1980), and it explores reasons for why the FEP may be a better option than the recently formulated framework of autopoietic enactivism (Di Paolo 2009; Thompson 2007). That said this critical examination does not imply that enactive approaches to mind (Chemero 2009; De Jaegher and Di Paolo 2007; Hutto and Myin 2013) and the FEP are mutually exclusive. Indeed, it is not my intention to claim that enactivism and the FEP cannot complement one another and so jointly combat orthodox 'cognitivist' view in cognitive science, viz., the computational and representational theory of mind.[1]

In this paper, the focus is restricted to the strong formulation of the life–mind continuity thesis. Before formulating my critical points, let me, as I did above, identify some of the implications of my overall claim. It will be argued that the FEP and the AT depart from one another in various ways. However, this does not preclude the AT from playing a part in a FEP explanation of the life–mind continuity thesis. Indeed, even if there are particular aspects of the AT that can be shown to be genuinely complementary to the FEP, there may still be good reasons to address the life–mind continuity thesis through the lens of the FEP. With these points of clarification out of the way, let us turn to consider the arguments to be developed.[2]

The first reason to prefer the FEP to the AT, as classically formulated by Maturana and Varela (1980), has to do with the nature of the causal-explanatory relation between living systems and their environment (Sect. 4). The autopoietic notion of life focuses on *self-production*, whereas the FEP emphasizes *self-preservation*. This difference is important. The notion of self-production highlights how processes of a system self-maintain *from within* the system itself. Thus for the AT, the explanatory relation between living systems and the environment takes an *internalist form*, reducing the role of the environment in homeostasis (the process whereby a system self-regulates its internal states given a fluctuating and unpredictable external environment). The causal-explanatory relation between living system and environment can also take an

[1] See Kirchhoff (2015a, b) for arguments that show how the FEP and enactivism can work together in complementary ways against cognitivist theories of mind.

[2] Given that the FEP and aspects of the AT can be shown to be complementary, would it not be better to unite these two frameworks and use them to jointly criticize view that object to a deep continuity between life and mind? This is certainly possible. But it is a task that will have to wait for another occasion.

externalist form.[3] In contrast to the AT, the FEP provides an account of the processes involved in self-maintenance that is premised on the environment of the living system, and the system's action on its environmental niche.[4] This is an example of an externalist causal-explanatory relation. This difference between the AT and the FEP is significant for unifying life and mind. Given that only externalism about the causal-explanatory relation between organism and environment accommodates adaptivity, and because the latter is necessary for mind, it will be argued that we have better reasons for preferring the FEP than the AT even though autopoiesis is a necessary component of life.

The second reason to prefer the FEP to the original formulation of the AT turns on considering an objection to the FEP (Sect. 5). This is the dark room problem. It states: if all organisms 'care' about is minimizing free energy, viz., minimizing the probability of entering into non-anticipated (surprising) states, then why would organisms not simply find a dark (non-surprising) room (environment) and stay there? If correct, the FEP should not be our preferred framework by which to address the life–mind continuity thesis given that it ultimately fails to explain phenotypic diversity. I will argue (i) that the AT is open to a variation of the dark room problem, and (ii) given the connection between variational free energy and thermodynamic free energy, the dark room objection is not a problem for the FEP but a problem for life.

In the final part of the paper, I extend my analysis to developments in the enactivist autopoietic literature (Di Paolo 2009; Froese and Stewart 2010; Thompson 2007). I start by establishing that the FEP has strong affinities with the framework of autopoietic enactivism. They both pursue a unified view of life and mind as rooted in adaptivity. Hence, the FEP and the autopoietic enactivist framework can be shown to be genuinely compatible, thus raising the possibility of a joint venture when it comes to answering the life–mind continuity thesis.[5] Yet the FEP and autopoietic enactivism are not identical. Thus, I end this paper by providing an admittedly brief comparison of their respective strengths and weaknesses—a full discussion of this issue will be a task for another occasion.

2 Keeping entropy in check: thermodynamic and variational free energy

Given that life exists, it is possible to infer that living systems make use of the second law of thermodynamics to remain far from thermodynamic equilibrium. The second law states that the entropy of non-open systems will increase unavoidably with time (Attard 2012). Thermodynamically, entropy is a measure of energetic disorder. Information theoretically, entropy is the average of uncertain (surprising) results sampled from inferences on a probability distribution. A sample with low entropy, for example, implies that the outcome is highly predictable, relative to the mean of the sample space (Friston 2010). Unlike closed-systems, living systems are dissipative, open-systems.

[3] The distinction between internalist and externalist forms of the causal-explanatory relation between organism and environment is due to Godfrey-Smith (1996).

[4] My account of the FEP is not comprehensive. See Friston (2005, 2009, 2013) and Friston and Stephan (2007), for detailed (mathematical) treatments.

[5] Thanks to an anonymous reviewer for pressing this point.

A dissipate systems is able to minimize entropy by exchanging energy with its milieu. The defining characteristic of living systems, then, is that such systems maintain homeostasis in the face of a constantly fluctuating environment (see e.g., Ashby 1962; Friston 2010; Haken 1983; Kauffman 1983; Nicolis and Prigogine 1977).

With this in mind, the next aim in this section is to illustrate two different conceptions of free energy by which to minimize entropy. This will set the stage for much of what is to follow in this paper.

Thermodynamic free energy is a measure of the energy that is available for a system to do useful work (McEvoy 2002). Hence, it is a measure of the average of energy that can be retrieved from a system and put to use. How should we understand this? In simplifying matters, let us first consider the case of the Physicist's Daughter. One day she arrives home, and asks: "Daddy, the school teacher told us we must conserve energy. But you told me energy is *always* conserved, no matter what you do" (Frautschi 1990, p. 12; italics in original). To see this, we can help ourselves to a different example: water behind a dam wall. In its initial state, water has potential energy. Yet as it starts to flow, its potential energy is converted into kinetic energy. When the water finally hits the pond below, its kinetic energy is transformed into heat and random turbulence. In this case, like the Physicist told his daughter, the energy is conserved throughout. The central issue is that while energy is conserved, it becomes more and more disorganized (Frautschi 1990). Since energy tends towards more disorganized forms, it becomes more and more difficult to extract it to do work. A technical term from statistical physics, which is a measure of energy available to do useful work, and which decreases with time, is *free energy*. A technical term of the measure of disorder, which increases with time, is *entropy*. It is in this precise sense that thermodynamic free energy amounts to the "difference between the energy and the entropy of a system" (Friston and Stephan 2007, p. 419).

This is thermodynamic free energy. The FEP makes use of variational free energy—which is different but related to thermodynamic free energy. Variational free energy originates in statistics, where it is understood as "a measure of statistical probability distributions" (Friston and Stephan 2007, p. 420). As we will see later, the long-term average of free energy is entropy. This means that minimizing free energy amounts to minimizing entropy. Furthermore, minimizing variational free energy is taken to be an inferential process, involving approximate Bayesian inference on probability distributions. The FEP applies this notion to explain how living systems minimize entropy in the face of changing environment (Friston 2012a). According to the FEP, living systems are able to keep an upper bound on entropy in virtue of minimizing variational free energy, where the latter amounts to generating inferences about environmental states. The implication is that living systems—to keep entropy within bounds—minimize free energy by predicting the probability of sensory input conditioned on their own internal models of the likelihood of encountering such-and-such sensory input given being in such-and-such situations (Friston and Stephan 2007; Hohwy 2015). The driving idea behind this proposal is that living systems become *models* of their local niche given that such systems distil, on average and over time, causal-statistical regularities of their niche and come to embody these regularities in their gross-bodily form and internal global dynamics (Friston 2012a, 2013). Given that the average of free energy (or surprise) is entropy, minimizing entropy is effec-

tively to maximize model evidence (Friston 2012a). This point is nicely put captured in the following quote by Friston, who says that living systems "become models of the causal structure in their local environment, enabling them to predict what will happen next and counter *surprising* violations of those predictions" (2012a, p. 2101; italics added). At this point, two things are important to be clear about.

First, some might say that talk about 'own internal models' is inconsistent with the claim that the FEP is committed to explanatory externalism about the organism-environment relation. For example, Froese and Ikegami (2013) charge the FEP with entailing a modern day version of internalism. Here care is needed in differentiating predictive coding from the FEP. The former is based in Helmholtz's theory of perception as unconscious inference (1860/1962). Hohwy (2014) has also argued that predictive coding is committed to internalism. But the FEP is not predictive coding. It is a much broader theoretical framework for addressing self-organization in biological systems, focusing on the interdependency of brain, body, and local environment (Bauson et al. 2005; Bruineberg and Rietveld 2014; Bruineberg et al., under review; Friston 2012b). I will say more about this in Sects. 4 and 5.

Second, free energy can be shown to bound surprise, where surprise is an information-theoretic measure of improbability, by minimizing the possibility of entering into surprising (non-anticipated and therefore improbable) states. Here a state has high surprise if it is deemed unlikely to occur relative to the embodied model of living systems—a system's expectations about the pattern of regularities in the world must likely to have caused its sensory input. Were a state to be consistently high in surprise it would be a state with high entropy. On the other hand, if a system is able to predict the external causes of its sensory input, then it will be in a state with low entropy (and hence, low surprise). Thus, living systems, by modeling their embedding niche, come to "acquire a homeostasis and can limit the number of states they find themselves in" (Friston 2012a, p. 2101) We can then say of the FEP that it illustrates how living systems are able to maintain themselves within a limited number of states—with low entropy—by combining probabilistic descriptions of internal systemic states (e.g. neural states) with approximate Bayesian inference and information theory (Friston 2005, 2012a, b).

What is the connection between thermodynamic free energy and variational free energy, where the latter is unpacked as minimizing free energy through the use of approximate Bayesian inference?

To see how variational free energy is connected to thermodynamic free energy, consider that because the latter is a bound of surprise—on average and over time—relative to a model, it is can also be considered as "*accuracy* minus *complexity*" (Hobson and Friston 2014, p. 19; italics in original). The basic idea is: if an inference on some probability distribution turns out to be accurate—if it has a 'grip' on the external patterns causing sensory input—then the prediction functions so as to minimize surprise. We can also reverse this by considering *complexity* minus *accuracy*. Were the brain, for example, not to receive sensory input, the issue of accuracy does not turn up for the simple reason that there is no flow of sensory input to determine the accuracy of. Yet, in cases of sensory deprivation, the brain can counter the increase of entropy by minimizing complexity (where complexity is understood as an increase in disorder). We can then explain the relation between these two kinds of free energy by appeal to the

idea that thermodynamic free energy is "based upon macroscopic processes pertaining to probability distributions over the microstates of a canonical ensemble [i.e. states within a system], while [variational free energy] pertains to probability distributions over hidden causes of sensory exchange with the external milieu" (Hobson and Friston 2014, p. 19). That is, it is possible to show that in the case of a lack of sensory stimuli, variational and thermodynamic free energy, share the same basic minimum (Sengupta et al. 2013).

3 Autopoiesis and the free energy principle: continuity and beyond

3.1 Organizational properties of living system

Is it possible to specify the organization of living systems? We shall consider the answer to this question through a comparison of the AT to the FEP. The starting point for these theories is that whatever organization a living system has its function must be to facilitate self-maintenance.

Before going into further details, a small philosophical aside is in order. That is, whatever the realization function of the process of self-maintenance is, it is not 'located' in the materialities that make up the system components. Thus, the organization of living systems proposed by the AT and the FEP is not bound, with the modal strength of logical necessity, to the biology of living systems. Proponents of the AT and the FEP are explicit about this. For the AT, Thompson puts this non-bio-chauvinistic point as follows, saying that "the autopoietic characterization of minimal life, though based on the living cell, sets out the general pattern of which DNA-based life is one possible expression" (2011, p. 102). Friston is equally clear, when he states that the FEP "applies to any … system that resists a tendency to disorder; from single-cell organisms to social networks" (2009, p. 293).

The AT explains this general pattern of which DNA-based life is one possibility, to use Thompson's phrase, by appeal to autopoiesis. A system is able to ensure low entropic levels in virtue of being able to *self-produce* over multiple time-scales. The FEP, as we have seen, states that minimizing free energy is equivalent to minimizing entropy on the assumption that entropy is the sum of surprise.

We can help understand what is involved in autopoiesis by considering the following passages by Varela and colleagues: "The bacterial cell is the simplest of living systems because it possesses the capacity to produce, through a network of chemical processes, all the chemical components which lead to the constitution of a distinct, bounded unit" (1997, p. 75). Maturana and Varela (1980) claim: "the notion of autopoiesis is necessary and sufficient to characterize the organization of living systems" (1980, p. 82). The necessity claim can be formulated along the following lines. For a system to be a living system it must, here-and-now, instantiate two properties: (i) it must be a reactive network—one capable of generating its own component processes; and (ii) it must establish a boundary separating itself from the environment. The sufficiency claim is as follows: for any system to instantiate these two properties guarantees that it is a living system. The two properties stand in a relationship of co-constitution. Consider, on the one hand, that it is the process of self-production that constitutes the system

and its boundaries. Yet, on the other, without creating its own boundaries, the system would not be able to self-produce. It would disperse into its surrounding medium.[6]

The FEP accommodates the requirement of AT. In this respect, the AT's focus on the process of autopoiesis the FEP. Indeed, one can show that the process of autopoiesis is a process that minimizes free energy given that for an organism to maintain a model of itself and its environment it must minimally self-produce the components required to carry out the process of preserving its generative model and Markov blanket (Friston 2013, p. 1) In this respect, the FEP and the AT converge on the following organizational property for living systems: self-maintenance through a process of autopoiesis, which results in homoeostasis.

Yet the FEP adds more to life than AT. To see this, consider again that surprise is the negative log probability of some state, i.e. that the surprise of being in such-and-such a state increases the more improbable it is for a system to be in such-and-such situations. Hence, entropy is the long run sum of surprise. Given that living systems do not have direct access to their local niche—to determine whether any state is surprising—it follows that such systems must embody a probabilistic model of itself and its environment from which the possible surprise of any given state can be measured. According to Friston and Stephan (2007), organisms can improve their probabilistic models, and so minimize free energy, by changing two quantities on which free energy depends: an organism can act in its environment, thereby changing the sensory input it receives; or an organism can change its model by changing its internal states. Hence, for the FEP, living systems minimize surprise on the basis of embodying a probabilistic model of themselves and their environment. This is a step beyond the appeal to mere AT in explaining the process of self-maintenance yet consistent with the idea that the "generative model thus functions—just as an enactivist might insist—to enable and maintain a structural coupling in which the viability of the organism is preserved" (Clark 2015, p. 22).[7]

3.2 Boundaries of living systems and their characteristics

Specifying the organization of living systems is among the key issues that any theory of life must minimally address. A different but related question that any theory of

[6] As Varela specifies: "An autopoietic system—the minimal living organization—is one that continuously produces the components that specify it, while at the same time realizing it (the system) as a concrete unity in space and time, which makes the network of production of components possible" (1997, p. 75).

[7] Some readers will no doubt make the objection that if A cannot directly access B, but must instead model B by sampling across probability distributions, then this commits the FEP to internalism. But this conclusion need not follow. Consider, if A stands in a relation to B, whatever that relation might be, then A and B are not one and the same but separated (in some fashion). In the literature on dynamical systems, it is common to treat two separate pendulums as *coupled*, and therefore as constituting a nonlinearly coupled dynamical systems—as a system comprising both A and B. However, even if A and B are coupled in this precise mathematical sense, there is still a schism between A and B. After all, A and B are not identical. But it is far from clear that such a separation entails commitment to any kind of internalism. In fact, it can be shown that any dynamical system A—e.g. an organism—coupled with a second dynamical system B—e.g. an environment—"can be said to 'infer' the 'hidden causes' of its 'input' (the dynamics of B) when it reliably covaries with the dynamics of B and it is robust to the noise inherent in the coupling" (Bruineberg and Rietveld 2014, p. 7).

life must take seriously is: Does the organization of living systems have a specific boundary? And what are the characteristics of such boundaries?

An autopoietic system is often taken to be an autonomous system. An autonomous system is a system that continuously regenerates and gives rise to the dynamic processes that produce such systems. Here is the crucial point. Autonomous systems have what is referred to as *operational closure*. Here 'closure' does not entail that such systems are somehow isolated or closed off from the environment. 'Closure' is technically *recursivity*, which is a mathematical concept. As Varela states: "It is exhypothesis evident that an autopoietic system depends on its physicochemical milieu for its conversation as a separate entity, otherwise it would dissolve back into it" (1997, p. 78). Recursivity, then, is defined as the circular and recursive network of processes that specify the system as a singular unity (Thompson 2007, pp. 44–45; Varela 1997, p. 82). Operational closure makes explicit that every process internal to the system is affected by and affects at least one other process in the system. In other words, "there are no processes that are not conditioned by other processes in the network" (Di Paolo 2009, pp. 15–16). It does not follow from this that processes outside the system do not have a causal impact on the system. Rather, it implies that environmental processes are not constituent parts of the operationally closed system (Di Paolo 2009, p. 16).

The FEP accommodates the requirement for operational closure. This is another way in which the two frameworks can be seen as complementary, and therefore not mutually exclusive. As we saw (in Sect. 2), the FEP conceives of living systems as *ergodic*. This means that on average there is a high probability that living systems will occupy a limited set of states. One modeling assumption is that if a system is ergodic then it can be considered as instantiating a *Markov* blanket, enabling it "to actively maintain its structural and dynamical integrity" (Friston 2013, p. 2; italics omitted). A Markov blanket constitutes a set of states that separates internal and environmental states. Markov blankets depend for their existence on *reciprocal coupling* between constituent processes. It can be shown that Markov blankets operate in much the same way as a cell boundary (Friston 2013). This is because the states comprising the blanket are functionally and dynamically dependent upon each other. Markov blankets can thus be understood as the statistical definition of operational closure, demarcating internal from external states, which, at the same time, establish the dynamics required to defining the system as a unity.

The FEP goes beyond operational closure. For the FEP, Markov blankets give rise to a partitioning of states into internal and external states. The external states are distinct from the internal states given the statistical blanket separating these states. The internal states can be partitioned further into states that are, and are not, children of external states. Here 'children of external states' are perceptual states, whereas 'children of internal states' are active (action) states. Given this, it is possible to partition states of a Markov blanket into internal, sensory, active, and external states. Friston explains the interdependency between the states in question as follows:

> External states cause sensory states that influence—but are not influenced by—internal states, while internal states cause active states that influence—but are not influenced by—external states. Crucially, the dependencies induced by Markov blankets create a *circular causality* that is reminiscent of *the action–perception*

> *cycle.* The circular causality here means that external states cause changes in internal states, via sensory states, while the internal states couple back to the external states through active states—such that *internal and external states cause each other in a reciprocal fashion.* This circular causality may be a fundamental and ubiquitous causal architecture for self-organization. (Friston 2013, pp. 2–3, italics added)[8]

It is this precise partitioning of states that explains why internal states must produce predictions about external states. They must do so to minimize (keep an upper bound on) surprise, thereby working to secure low entropy. Any ergodic system can thus be seen to possess a Markov blanket upon which its probabilistic models are conditioned. This is a step beyond the appeal to operational closure in explaining the boundaries of self-maintaining systems.

I now turn to developing my arguments outlined above.

4 Argument #1: two kinds of causal, explanatory relations

The first argument turns on the following distinction between *self-production* and *self-preservation*. This distinction maps onto two different kinds of causal-explanatory relations between organism and environment: an *internalist* kind, and an *externalist* kind. As we have seen, AT focuses on self-production. Thus, it follows that AT is based on a causal, explanatory relationship between the organism and the environment of an internalist kind. To see this, consider again the notion of operational closure. Operational closure is what establishes a system as a unity. At the same time it specifies what is exterior to the system. Operational closure enforces what is external to any given autopoietic system, and what turns out to be external can only be understood *from the inside of a system* (Varela 1997, p. 78). Moreover, the activity that leads to life is conceived of from within the system. Life as we know it, for the AT, is "dictated by the internal system's organization" (Bitbol and Luisi 2004, p. 99).

To avoid confusion, when I say that the AT takes an internalist form of the causal-explanatory relation between organism and environment, I mean it in precisely the way expressed by Godfrey-Smith (1996): "it can also appear in a more internalist form, as exemplified by the 'autopoietic' conception of life given by Maturana and Varela (1980)" (1996, p. 318). What I am *not* saying is that the AT is committed to *metaphysical internalism* when it comes to expressing the relationship between mind and its material realizers. The point is merely that when explaining the organization of a living system, the classical formulation of the AT focuses on dynamics *internal* to a system. I am not the first to make this point. Godfrey-Smith makes it in his (1996) article. Chemero (2012), an enactivist, makes a similar observation, as he says:

[8] One might claim that the FEP overplays its capacity to explain the interdependence between sensation and movement. After all, this is not special to the FEP but a baseline condition of any dynamical system that is coupled to its environment. This is observation is correct. But it is not a problem for the FEP given that it based in dynamical systems theory and statistical physics.

> [P]roponents of autopoiesis do not claim that the autopoietic systems are entirely walled off from the environment. For one thing, proponents of autopoiesis claim that autopoietic systems are "structurally coupled" to the environment. For another, Maturana and Varela (1980) would never deny that autopoietic systems cause changes to the environment; they just think that those changes are by-products and are not important for describing the nature of the system. (2012, pp. 53–54)

The claim that the AT takes an internalist form of the causal-explanatory relation between organism and environment is compatible with the enactivist claim that minds are realizes in distributed networks encompassing neural and non-neural features. So the point is merely that once one adopts an internalist perspective on this particular causal-explanatory relation, it reduces the role of the environment itself in explaining a system's integrity over time.

Prima facie, at least, this sets up an interesting contrast with the FEP. The FEP emphasizes self-preservation. If I am right to say that self-production maps onto explanatory internalism, and that self-preservation maps onto explanatory externalism, then the FEP is premised on a causal-explanatory relationship between organism and environment that is of an externalist kind. In contrast to the AT, the FEP is conditioned on the idea that a living system's "structural and functional organization is maintained by causal structure in the environment" (Friston and Stephan 2007, p. 418). If this is correct, it is then possible to show that the reason for why an organism is able to minimize free energy by predicting future states is due to the fact that "the hierarchical [statistical] structure of our brains is transcribed from causal [statistical] hierarchies in the environment" (Friston and Stephan 2007, p. 418). Given that the environment is non-stationary, organisms can optimize their predictions through embodied activity in the environment—with subsequent change and approximate optimization of their generative model and Markov blanket. It is because organisms can act on the environment that they can occupy a domain far from (terminal) phase-boundaries. Indeed, organisms strive to achieve a maximal fit between their probabilistic models and environmental niche via embodied activity (Friston et al. 2012; see also Bruineberg and Rietveld 2014). This is also known as *active inference* (Friston et al. 2011). It is one of the means by which organisms can minimize free energy. The basic idea is that action minimizes surprise by enacting a shift in the environment. Action is 'in' perception—to use an enactivist phrase (Noë 2004)—enabling an organism to achieve a tight grip on the world by producing sensory input—via embodied activity—that align with its predictions about external causes of the sensory signals received (Friston 2009).

We can help understand this emphasis on embodied activity further by considering this imaginative example by Friston and Stephan (2007) (Fig. 1).

The contrast is between a non-biological self-organizing system (a normal snowflake) and one fitted with wings, enabling it to engage in active inference. Only the special snowflake endowed with wings is able to act on the environment. This opens a space for it to remain far removed from its phase-boundaries. The phase-boundary is here depicted as temperature. The normal snowflake, by contrast, cannot act on its niche, and will inevitably fall and meet its phase-boundary, i.e. it will melt because of an increase in temperature. The key "difference between the normal snowflake and

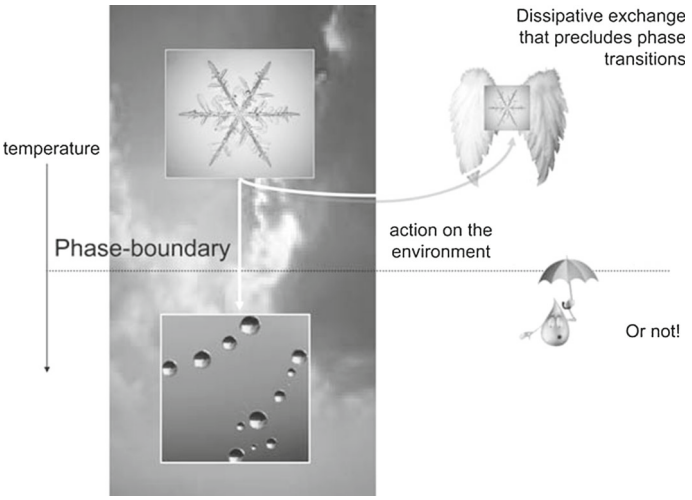

Fig. 1 Illustration of the difference between mere self-organizing systems and adaptive self-organizing systems (Friston and Stephan 2007, p. 423; used with kind permission from Karl Friston and the Welcome Trust)

the adaptive snowflake is the ability to change their relationship with the environment and maintain thermodynamic homeostasis" (Friston and Stephan 2007, p. 423). This example highlights a key feature of life. To persist organisms must avoid certain phase-transitions. And to avoid those, organisms must not only adapt to their milieu but must also adapt their environment to their own needs (Bauson et al. 2005).[9] Hence, *adaptivity* is a necessary feature of life.[10]

This is significant for the life–mind continuity thesis. First, cognition is usually taken to involve a future-oriented regulation (Bitbol and Luisi 2004; Clark 1997; Di Paolo 2009; Varela et al. 1991). It follows that cognition is *prima facie* a relational feature, enabling living systems to improve conditions of self-maintenance through *adaptive* behavior in their environmental niche (Clark 2001; Hendricks-Jansen 1996; Hutto and Myin 2013). Second, the life–mind continuity thesis states that cognition is foreshadowed in life. This means that the relational, future-oriented aspect of cognition must be explicitly defined as part of life. If not, then cognition and life would not share the same basic set of properties.

The problem with the AT is then very clear. As we have seen, the classical formulation of the AT is consistent with but downplays adaptivity given its internalist perspective on the causal-explanatory relationship between organism and environ-

[9] The FEP is not committed to the idea that *all* phase transitions must be avoided. Indeed, there is no need to think that the FEP is incompatible with phase transitions such as those involved in the transitions from egg-caterpillar-cocoon-butterfly. Thanks to an anonymous reviewer for mentioning this example.

[10] One might press the claim that the FEP is premised on adaptivity. But this seems difficult to do given that a key assumption of the FEP is that without active inference (embodied activity) organisms would not be able to minimize free energy (Hohwy 2012). The reason for this is that active inference—on the FEP formulation—is a precondition for minimizing free energy under hierarchical generative models and Markov blankets.

ment. While this formulation of the AT is consistent with a future-oriented notion of cognition, it is not one that is emphasized by the notion of autopoiesis. That is a problem if the AT intends to unite life and cognition. Since cognition is not foreshadowed in the properties of autopoiesis, then autopoiesis is not sufficient for cognition, even if it guarantees life. Bitbol and Luisi (2004) make a similar observation. Yet they do not arrive at this point by reference to explanatory internalism, as I have done here. Nevertheless, they go on to say:

> [A]utopoiesis could at least include the necessity of cognition-like relations for its own maintenance in its definition. Conversely, if … this is not done, in other words if cognition (a) remains excluded from the definition of autopoiesis (which focuses on internal organization and self-generation) and (b) is nevertheless construed as indispensable for life, then autopoiesis and cognition are distinct processes. (2004, p. 100)

If this is true, one might object that the AT is unable to adequately address the life–mind continuity thesis. Di Paolo says precisely this about the unifying and explanatory power of the AT, as classically formulated:

> We simply cannot derive from the definition of autopoiesis that an autopoietic system will attempt to ameliorate a situation that leads otherwise to the future loss of autopoiesis … Improving the conditions of self-production is not part of the definition of autopoiesis, nor can it be deduced from this definition which simply states that the system self-produces. Hence, an autopoietic system that is able to operate according to potential future states cannot do so just because it is autopoietic. [Hence], it would seem that classical autopoiesis on its own and cognition cannot travel together from this point onwards. (2009, p. 13)[11]

Indeed, when considering the question of whether cognition—or sense-making—is constitutively dependent on autopoiesis, Thompson admits, "to being unsure" (2011, p. 40) for these particular reasons:

> [A]s work on autonomous systems in AI and robotics suggests (see Barandiaran et al. 2009; Froese and Ziemke 2009), it seems conceivable that there could be an adaptive self-constituting system that was not based on autopoietic constituents. For example, perhaps it is possible to bypass autopoiesis and construct directly a sensorimotor agent that achieves its autonomy at the level of an adaptive and organizationally closed sensorimotor loop. (2011, p. 40)

What does this entail? We already know that autopoiesis per se does not guarantee the existence of minds (or sense-making). If it is conceivable that an adaptive sense-making system is not constitutively dependent on autopoiesis, then it follows that mere autopoiesis is neither sufficient nor necessary for mind.

Although the AT is attractive when it comes to addressing life, appeals to nothing but autopoiesis face immediate difficulties when having to explain the relationship between life and cognition. Thus, even if AT guarantees life, it does not guarantee cog-

[11] I consider Di Paolo's (2009) alternative to mere autopoiesis in Sect. 6.

nition. In other words, even if autopoiesis is necessary and sufficient for life (though some contest this), it does not entail that autopoiesis is sufficient for cognition. Consequently, the AT falls short of accommodating the strong life–mind continuity thesis.

The FEP is not open to this charge. It incorporates *adaptivity*—viz., the future-oriented aspect of cognition—from the very beginning. This can be shown by the notion of a Markov blanket. Markov blankets are (in part) constituted by dependencies between processes on the inside of the blanket. Still, internal states can change external states via active inference, thereby allowing organisms to remain (for some timescale) removed from terminal phase-boundaries. This is what Friston and Stephan (2007) had in mind, as they say: "systems that minimize their free energy respond to environmental changes *adaptively*" (2007, p. 428; italics in original). If correct, the FEP is based on the premise that for systems to minimize free energy they must necessarily regulate their activity in a future-oriented fashion. In fact, a core feature of minimizing free energy is based on dealing with uncertainty about the environment. One modeling assumption here is that this implies sampling the environment in an epistemic sense, seeking a solution to uncertainty and therefore surprise in the future. Examples of this include saccadic eye movement, explorative behavior, and so on. Here life and cognition are unified in that both life and cognition share the same basic future-oriented, adaptive property.

This is the missing link in the AT. Therefore, we can conclude, the FEP is better able than the AT to address the life–mind continuity thesis. It is a view of the mind-life relationship that implies a form of *biopsychism*: that to live is to be cognitive. It is not a position entailed by mere autopoiesis.

5 Argument #2: stay out of the dark room if you want to survive

Consider again the FEP. The whole function of the brain, including an organism's phenotypic traits, is free energy minimization. As Clark puts it, this sole attention to minimizing free energy "captures something crucial about the way that spending metabolic money to build complex brains [and phenotypic traits] pays dividends in the search for adaptive success" (2013, p. 181). The claim is clear: minimization of free energy fuels and explains adaptive success.

Despite this emphasis on explaining adaptivity, there is an objection to precisely this point. It is the dark room problem. It goes like this. If the sole function of an organism, including but only limited to its brain, is to reduce surprise, then organisms should—in virtue of this imperative—occupy states with the lowest possible levels of surprise. Where else could that be than in the least stimulating environment? That is, if all an organism cares about is minimizing surprise, then it should take up position in the nearest dark room, or concealed situation, and stay there. Surely this would be the best way to ensure low levels of surprise. If correct, the FEP is not the best framework for explaining what fuels and augments adaptive success. For it fails to account for what drives cognitive, behavioral, and phenotypic development.

If this problem is taken to imply skepticism about the FEP's ability to explain how complex traits arise, then this kind of skepticism applies equally to the AT. That is, it is not possible to derive facts about the development of traits from facts about mere autopoiesis. This follows from the fact that it is not possible to derive facts about

adaptivity from facts about autopoiesis. In fact, Varela comes close to admitting this, as he says:

> It is important to point out two aspects of the living phenomenology that this autopoietic characterization does not address … how this basic autopoietic organization, present at the origin of terrestrial life, becomes progressively complexified through reproductive mechanisms … give rise to the variety of pro- and eukaryotic life on Earth today. (1997, p. 76)

Considered as an objection to the possibility of explaining the development of traits, the objection hits the AT as well. I now turn to consider whether the dark room problem is an actual problem for the FEP. We can start by noting that the dark room problem is a thought-experiment and that it can be shown to be not plausible as an objection to the FEP along a number of different dimensions.

First, as we have seen, minimizing free energy is also to minimize entropy on average and over time. In so doing, systems that minimize entropy by minimizing free energy are also minimizing uncertainty—relative to their generative model—of about the environment. Note that there is a recurrent and reciprocal dynamics between organism and environment, where organisms strive towards optimizing their grip (its generative model) on the niche, all the while the niche provides evidence—of lack of evidence—for the generative model. If one were to remove the environment, concealing an organism in a black box, there would be nothing for the organism to model. This is a problem for the dark room objection. For, according to the FEP, an organism depends for its existence on modeling its environment. Hence, secluding an organism in a dark room by taking away its environment results in the organism ceasing to exist.

Second, what would the FEP predict if one concealed an organism in a black box? As we have seen, in the absence of sensory stimuli, organisms will not be able to work towards making their generative models more accurate. Hence, the only means by which an organism could minimize free energy is by minimizing complexity (Hobson and Friston (2014). This implies that the only thing that the brain can minimize is thermodynamic free energy, leading (inevitably) to thermodynamic equilibrium. According to the FEP, the latter implies death. This is what the FEP predicts. To survive, organisms must minimize complexity as it nears thermodynamic equilibrium. Yet, living systems would not be able to minimize complexity over substantial periods of time given deprivation of energetic stimuli from the environment. To illustrate the non-trivial role of the environment in driving homeostasis, we can add to the case of 'water behind a dam wall' (Sect. 2). We can ask: how can we get the water back up into the reservoir to do more work? The answer is: the sun does this for us. As Frautschi reminds us: "Sunlight provides energy, some of which evaporates water and raises it up to the clouds, whence it finds its way into the reservoir ready to do more work" (1990, p. 12). Moreover, according to Frautschi, "sunlight comes in an organized form, and weather systems and plants and animals all tap some of this organized energy to maintain and develop their own organization, thereby avoiding the tendency toward equilibrium" (1990, p. 12) Depriving organisms of sensory input, concealing them in a dark room, leads to organisms loosing the structure of their generative model and

Markov blanket, and will therefore die. If correct, the dark room problem fails to be a worry for the FEP.

There is a residual worry here. Even if the dark room problem is not a problem for the FEP, one might still argue that the FEP cannot explain the development of phenotypic traits. This worry also confronts the AT. Specifically, the worry is that the FEP is not able to explain the development of phenotypic traits, because it is unable to explain what such development ultimately depends on: learning-driven plasticity in a scaffolded learning environment. Menary (2015) has recently raised such a worry about the explanatory power (or lack thereof) of the FEP.

This worry is less strong than it initially appears. To see this, consider an analogy provided by Hohwy (2015): "The heart pumps blood, and this function is realized in part by the way the contraction of the heart muscle occurs—a process that depends on intricate ion flows across heart cell membranes" (2015, p. 2). It would be wrong to say that the function of the heart is not to pump blood by pointing out that what really happens in the heart are complex forms of cellular ion flow. One can certainly learn more about the heart by considering cellular and molecular aspects about it. But it does not follow from such considerations that cellular and molecular features are not what in part realizes the heart's primary function: to pump blood. The same can be said about the brain. According to Hohwy: "one cannot object to the free energy principle by pointing to facts about what the brain does (e.g., what happens as action potentials are generated)" (2015, p. 2). The reason for this is the same as the one concerning the heart. That is, neural plasticity and neural reuse might—just like ion flows across heart cell membranes—be part of what realizes free energy minimization. If this is correct, it casts a new light on Menary's claim. It need not be viewed as an objection to the FEP. Instead it can more fruitfully be considered as a call for further explanatory work such as how neural plasticity and neural redeployment can be realizations of free energy minimization. Indeed, it is not difficult to establish that associative plasticity and Hebbian learning are processes which themselves minimize free energy under generative models. Hence, we have no good reasons to infer from the many variations of the dark room problem to the falsehood of the FEP. Instead, it is because the FEP provides a unifying model for explaining the function of the brain, coupled to the rest of the organism, that is has explanatory power in the first place.

6 Free energy and autopoietic enactivism

At this stage you will no doubt be eager to make the observation that I have left out a significant development in the autopoietic literature. I have omitted, you will insist, that AT has recently been augmented with adaptivity. If so, you might not accept my argument. And your reason will be that I have failed to consider what Di Paolo calls *autopoiesis+* (2009, p. 13), or something close to it. This point is entirely correct. Yet I have done so deliberately to consider it now. I start by establishing how these two approaches can be seen to come together.

For Di Paolo (2005, 2009), mere autopoiesis needs supplementation. It is not enough that a system is autopoietic for it to be cognitive (see also Thompson 2007). It is on this point that the FEP and autopoietic enactivism come together. To see this,

consider that both the FEP and autopoietic enactivism conceive of minds as established in the self-organizing and self-generating activities of living organisms. Indeed, for Di Paolo (2009), for a system to be cognitive it must not only be autopoietic it must also be adaptive. As we have seen, in the context of active inference, the FEP shares the idea that cognition is rooted in adaptive activities of organisms. Di Paolo defines adaptivity as a system's capacity to regulate its own states and its relation to its environmental niche with the result that, "if the states are sufficiently close to the limits of viability" (2009, p. 14), then: (i) these will be acted upon to recede from such limits, with the result that (ii) such states will be transformed into tendencies to prevent entering into states that are close to the limits of viability.

No doubt there is something to this addition. It is precisely why I argued, in the first place, that the FEP outcompetes the AT, as classically formulated by Maturana and Varela (1980). Of significance is that adaptivity is taken to be a complexification of AT (Thompson 2011). On one reading, this suggests that there are degrees of autopoiesis—hence the reference to autopoiesis^{+} (Di Paolo 2009). This shoehorns nicely with the FEP. Given that the FEP can be shown to apply to systems that are, at least arguably, not cognitive, it follows that mentality arises when organisms minimize free energy to a certain degree—viz., in the context of active inference. Indeed, for an open system to minimize free energy, it must act on its embedded environment. It must do so because for open systems to minimize free energy they must minimize uncertainty about the environment (see Friston et al. 2015b for a discussion of the equivalence between minimizing free energy and informational gain). The FEP, like the notion of autopoiesis^{+}, thus appears to be enough to account for adaptivity.[12] Moreover, if the future-oriented aspect of adaptivity is a form of cognition, then the FEP—like autopoiesis^{+}—is enough to guarantee cognition. Thus, on both accounts, life and mind share the same basic property: adaptivity—a capacity for future-oriented regulation of functional and structural integrity.

At this stage, proponents of autopoietic enactivism might raise the following worry: nothing in the FEP is able to account for the constitution of a meaningful perspective. This is a variation on the strong life–mind continuity thesis that I have been considering so far. Thompson (2011) calls it the *existential-phenomenological* sense of the life–mind continuity thesis, which he ascribes to Jonas (1966, 1968). The idea is that subjectivity—or an existential concern with staying alive—is conditioned on metabolic processes. As Thompson says: "Jonas argues that without metabolic self-construction, there would be no such thing as the constitution of a meaningful perspective by a system for that system. This idea provides the existentialist side of the deep continuity thesis" (2011, p. 41).

Let us grant that such an existential concern with life is premised on metabolic processes, and ask: is this beyond the reach of the FEP? It does not appear so. To see this, consider that the particular generative models that organisms come to exhibit of themselves and their niche is constrained by resource and complexity costs (FitzGerald et al. 2014). The imperative is: the more models an organism must generate the higher the complexity. As FitzGerald et al. note: "This fits comfortably with minimizing vari-

[12] As Friston and Stephan put it: "This principle … is sufficient to account for adaptive exchange with the environment by ensuring a bound on adaptive value is optimized" (2007, p. 427).

ational free energy—that necessarily entails a minimization of complexity" (2014, p. 4). Consider, also, that neuronal activity—which is what (in part) instantiates the models in questions—is costly in terms of metabolism (Laughlin et al. 1998). The FEP implies that if minimizing free energy is equivalent to minimizing the complexity in realizing generative models, and if high complexity corresponds to high metabolic activity, then to minimize free energy is also to minimize metabolic costs (FitzGerald et al. 2014). The existential-phenomenological interpretation of the life–mind continuity thesis takes metabolism as sufficient for instantiating a meaningful perspective—a concern with being and staying alive. We have just seen that the FEP accommodates metabolic activity. Thus, if metabolic activity is sufficient to ensure the existentialist-phenomenological life–mind continuity thesis, then the FEP is sufficient for handling the existentialist-phenomenological side of the life–mind continuity thesis.

The FEP and autopoietic enactivism thus depict the organism and the organismic niche as coupled together "in a process of mutual specification in which the simplest approximations apt to support a history of viable interaction are the ones that are learnt, selected, and maintained" (Clark 2015, p. 19).

Yet they also differ from one another. On the one hand, for autopoietic enactivism, the notion of 'sense-making' is central to an explanation of life–mind continuity (Di Paolo 2005, 2009; Froese and Di Paolo 2009; Weber and Varela 2002). It is defined as follows, by Di Paolo: "the instauration of a natural perspective from which encounters in the world are intrinsically meaningful for the organism following the norm established by the continuing process of self-production" (2005, pp. 429–430). On the other hand, the FEP is just more minimization of free energy.

A suspicion one might harbor is that if the FEP gives up on sense-making, then how can it explain what is central to life and mind? It is important to forestall a possible confusion here. The claim that the FEP turns on minimization of free energy need not imply a denial of sense-making. As Clark (2015) points out, the constructive aspect of the FEP "corresponds rather closely to the [enactivist] notion of 'enacting a world"' (2015, p. 18). Specifically, active inference puts organisms in a place to actively construct what they perceive via embodied activity in the world. Indeed, as Clark further states: "In this way, different organisms and individuals may selectively sample in ways that both actively construct and continuously confirm the existence of different 'worlds'. It is in this sense that, as Friston, Adams, and Montague (2012, p. 22) comment, our implicit and explicit models might be said to 'create their own data"' (2015, p. 20). So, by emphasizing free energy minimization all of the way up and all the way down, the FEP need not be cast as giving up on the notion of sense-making.

Although the notion of sense-making is attractive, it also faces problems. That is, the idea that sense-making is the right way to go in cognitive science is not obvious. For example, radical formulations of enactivism are skeptical of sense-making (Hutto and Myin 2013). These authors think that sense-making implies a view of basic minds "that create, carry, and consume [contentful] meanings" (Hutto and Myin 2013, p. 34). I think that there is something right about this skepticism. But the justification of this turns on whether sense-making is given an inflationary or a deflationary interpretation. Consider, as a case in point, that Di Paolo claims that sense-making requires an organism to "*evaluate* the situation" (Di Paolo 2009, p. 14; italics in original). An inflationary reading of this claim would seem to imply that sense-making is a

process of formulating contentful mental states and based on such states come to certain conclusions about the world. If, however, sense-making is given a deflationary articulation, which, I think, is what autopoietic enactivist intend to give in the first place, then it is far from obvious that talk of sense-making is inconsistent with talk of free energy minimization. To see this, consider that the FEPs talk of models and inference need not commit the FEP to positing either content or the existence of inner representations. Clark (2016) seems to agree, as he says:

> Whatever the use of the terms 'infer' and 'model' mean in these … free-energy minimisation accounts, they [do] not seem to imply the presence of inner models or content-bearing states of the kind imagined in traditional cognitive science. Instead, what are picked out seem to be physical processes defined over states that do not bear contents at all—neither richly reconstructive *nor* of any more 'action-oriented' kind. (2016, p. 18)

So, if the notion of sense-making entails the existence of meaningful content, then it is not at all clear that the FEP and autopoietic enactivism can be brought into a truly fruitful interplay—jointly pursing an account of the life–mind continuity thesis. Yet, if it can be shown that sense-making—like inference and anticipation—need not be cast in contentful terms—such as when the wavelength selectivity of photoreceptors reflects assumptions about the wavelength of ambient light (Friston et al. 2012, p. 5)—then this raises the possibility of a joint venture when it comes to answering the life–mind continuity thesis.[13]

7 Conclusion

In this paper, I have argued that there are reasons to prefer the FEP to the AT, as classically formulated, when considering the life–mind continuity thesis even if certain features of the AT can be shown to be complementary to the FEP. I have presented two arguments for why this is so. The first argument turned on a distinction between self-production and self-preservation. This distinction, I argued, maps onto a distinction between internalist and externalist forms of addressing the causal-explanatory relation between organism and environment. Given that only an externalist perspective on this relation can encompass adaptivity, I inferred to the claim that the FEP—but not the AT—has the resources needed to address the life–mind continuity thesis. The second argument considered why the FEP is immune to the dark room problem, whereas the theory of AT is open to a variation of this particular problem. I then argued, finally, that the FEP and recent work on autopoietic enactivism come together on the idea that cognition is premised on life and adaptivity, which opens a possible space for both proposals to work in unison when addressing the unity of life and mind.

Acknowledgements I would like to thank Julian Kiverstein, Eric Rietveld and Jelle Bruineberg for helpful comments on this paper. I am also grateful to all the participants at the Predictive Brain and Embodied,

[13] Although developing an answer to this issue is of significance, it is a task that will have to wait for another occasion.

Enactive Cognition workshop at the University of Wollongong. Finally I would like to thank two anonymous reviewers for highly constructive comments.

References

Ashby, W. R. (1962). Principles of the self-organising systems. In H. V. Foerster & G. W. Zopf Jr. (Eds.), *Principles of self-organization: Transactions of the University of Illinois symposium* (pp. 255–278). London: Pergamon Press.

Attard, P. (2012). *Non-equilibrium thermodynamics and statistical mechanics: Foundations and applications*. Oxford: Oxford University Press.

Barandiaran, X., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry and spatio-temporality in action. *Adaptive Behavior*, *17*, 367–386.

Bauson, G., Bergfeldt, N., & Ziemke, T. (2005). Brains, bodies, and beyond: Competitive co-evolution of robot controllers, morphologies and environments. *Genetic Programming and Evolvable Machines*, *6*, 25–51.

Bitbol, M., & Luisi, P. L. (2004). Autopoiesis with or without cognition: Defining life at its edge. *Journal of Royal Society Interface*, *1*, 99–107.

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, *8*, 1–14. doi:10.3389/fnhum.2014.00599.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: The MIT Press.

Chemero, A. (2012). Modeling self-organization with nonwellfounded set theory. *Ecological Psychology*, *24*(1), 46–59.

Clark, A. (1997). *Being there*. Cambridge, MA: MIT.

Clark, A. (2001). *Mindware*. Oxford: Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–253.

Clark, A. (2015). Radical predictive processing. *Southern Journal of Philosophy*, *53*, 1–25.

Clark, A. (2016). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Nous*. doi:10.1111/nous.12140.

De Jaegher, H., & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences*, *6*(4), 485–507.

Di Paolo, E. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences*, *4*, 97–125.

Di Paolo, E. (2009). Extended life. *Topoi*, *28*, 9–21.

FitzGerald, T., Dolan, R., & Friston, K. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, *8*, 1–11.

Frautschi, S. (1990). Entropy in an expanding universe. In B. Weber, D. Depew, & J. Smith (Eds.), *Entropy, information, and evolution* (pp. 11–22). Cambridge, MA: MIT.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*, 815–836.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. doi:10.1016/j.tics.2009.04.005.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. (2011). Embodied inference: Or 'i think therefore i am, if i am what i think'. In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment (cognition and communication)* (pp. 89–125). Exeter: Imprint Academic.

Friston, K. (2012a). A free energy principle for biological systems. *Entropy*, *14*, 2100–2121.

Friston, K. (2012b). Free energy and global dynamics. In M. Rabinovich, K. Friston, & P. Varona (Eds.), *Principles of brain dynamics: Global state interactions* (pp. 261–292). Cambridge, MA: MIT.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*, 20130475. doi:10.1098/rsif.2013.0475.

Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015a). Knowing one's place: A free energy approach to pattern regulation. *Journal of the Royal Society Interface*, *12*, 20141383. doi:10.1098/rsif.2014.1382.

Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104*, 137–160.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015b). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214.
Friston, K., & Stephan, K. E. (2007). Free energy and the brain. *Synthese*, *159*, 417–458.
Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3*(130), 1–7.
Froese, T., & Di Paolo, E. (2009). Sociality and the life-mind continuity thesis. *Phenomenology and the Cognitive Sciences*, *8*, 439–463.
Froese, T., & Ikegami, T. (2013). The brain is not an isolated "black box," nor is its goal to become one. *Behavioral and Brain Sciences*, *36*(3), 213–214.
Froese, T., & Stewart, J. (2010). Life after ashby: Ultrastability and the autopoietic foundations of biological individuality. *Cybernetics & Human Knowing*, *17*(4), 83–106.
Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, *173*, 466–500.
Godfrey-Smith, P. (1996). Spencer and dewey on life and mind. In M. Boden (Ed.), *The philosophy of artificial life* (pp. 314–331). Oxford: Oxford University Press.
Haken, H. (1983). *Synergistics: An Introduction. Non-equilibrium phase transitions and self-organisation in physics, chemistry and biology*. New York: Springer.
Helmholtz, H. (1860/1962). Handbuch der physiologischen optik (J. P. C. Southall, Ed., English Trans.) (Vol. 3). New York: Dover.
Hendricks-Jansen, H. (1996). In praise of interactive emergence, or why explanations don't have to wait for implementation. In M. Boden (Ed.), *The philosophy of artificial life*. Oxford: Oxford University Press.
Hobson, A., & Friston, K. (2014). Consciousness, dreams, and inference. *Journal of Consciousness Studies*, *21*(1–2), 6–32.
Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 1–14. doi:10.3389/fpsyg.2012.00096.
Hohwy, J. (2014). The self-evidencing brain. *Nous*, 1–27. doi:10.1111/nous.12062.
Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. Windt (Eds.), *Open MIND: 19(T)*. Frankfurt am Main: MIND Group. doi:10.15502/9783958570016.
Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT.
Jonas, H. (1966). *The phenomenon of life: Toward a philosophical biology*. Chicago: University of Chicago Press.
Jonas, H. (1968). Biological foundations of individuality. *International Philosophical Quarterly*, *8*, 231–251.
Kauffman, S. (1983). *The origins of order: Self-organization and selection in evolution*. Oxford: Oxford University Press.
Kirchhoff, M. D. (2015a). Experiential fantasies, prediction, and enactive minds. *Journal of Consciousness Studies*, *22*(3–4), 68–92.
Kirchhoff, M. D. (2015b). Species of realization and the free energy principle. *Australasian Journal of Philosophy*, *93*(4), 706–723.
Laughlin, S., van Steveninck, R., & Anderson, J. (1998). The metabolic cost of neural information. *Nature Neuroscience*, *1*, 36–41.
Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: D. Reidel.
McEvoy, P. (2002). *Classic theory: The theory of interacting systems*. San Francisco: Microanalytix.
Menary, R. (2015) What? Now. Predictive coding and enculturation—A reply to Regina E. Fabry. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 25(R)*. Frankfurt am Main: MIND Group. doi:10.15502/9783958571198
Nicolis, G., & Prigogine, I. (1977). *Self-organisation in non-equilibrium systems*. New York: Wiley.
Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT.
Sengupta, B., Stemmler, M. B., & Friston, K. (2013). Information and efficiency in the nervous system–A synthesis. *PLOS Computational Biology*, *9*(7), 1–12.
Stewart, J. (1996). Cognition = life: Implications for higher-level cognition. *Behavioral Processes*, *35*, 311–326.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
Thompson, E. (2011). Reply to commentaries. *Journal of Consciousness Studies*, *18*, 176–223.
Varela, F. (1997). Patterns of life: Intertwining identity and cognition. *Brain and Cognition*, *34*, 72–87.
Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT.
Weber, A., & Varela, F. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*, *1*, 97–125.