Biology & Philosophy (2021) 36:10
https://doi.org/10.1007/s10539-021-09787-1

The free energy principle: it's not about what it takes, it's about what took you there

Axel Constant[1]

Received: 12 March 2020 / Accepted: 8 February 2021 / Published online: 22 February 2021

Abstract

Philosophical writings on the free energy principle in the life sciences often give the impression that minimising free energy is sufficient for life. But minimising free energy is not a sufficient condition for life. In fact, one can perfectly well conceive of a system that actively minimises its free energy, and for this very reason moves inexorably towards death. So, where does the assumption of this entailment relation come from? There is indeed an entailment relation, but it goes the other way around: life entails minimising free energy. Put another way, if you exist, now, under the right conditions, it is because you've done something like minimising your free energy. However, the question of whether you will exist tomorrow cannot be settled purely by resorting to the fact that you will minimise your free energy to get there. The simple point I make in this paper is that the free energy principle is not concerned with the sufficient conditions of existence, but rather with what must have been the case, given that you exist. It's not about figuring out what it takes to be alive; it's about figuring out what took you there.

Keywords Free Energy Principle · Postdiction · Prediction · Historical sciences

Introduction

Sometimes, arguments in the literature on the Free Energy Principle (henceforth FEP) give the impression that in order to be alive, viz. to count as a living system, one must minimise free energy. Such a claim does not straightforwardly apply to the free energy principle, however, and this is what this paper will demonstrate. Minimising free energy does not entail life. Rather, the argument is that if you are alive, it probably means that you have done something like minimising your free energy, which is the (Bayes) optimal thing to do when your life depends upon solving

✉ Axel Constant
axel.constant.pruvost@gmail.com

[1] Theory and Method in Biosciences, Level 6, Charles Perkins Centre, The University of Sydney, D17, Johns Hopkins Drive (off Missenden Road), Sydney, NSW 2006, Australia

complex inference problems. This is a subtle, but crucial point to getting the story straight. I shall call this the 'entailment problem'; that is, the confusion in the entailment relation between free energy minimisation and life. Here, the notion of entailment refers to the implication (i.e., first order logical property) between free energy minimisation and the fact of 'displaying some life related processes'.

The entailment problem, it seems to me, stems from the fact that there are at least two types of claims one can conceive of when thinking about the relation between life and free energy minimisation. Or rather, about *survival*, and free energy minimisation; although, under the FEP, these appear to be synonymous. Minimising free energy is the process whereby one maintains one's structural integrity in face of environmental perturbations by revisiting one's most probable organisation of physiological states (Friston 2013; Kirchhoff 2015). It is in that sense that minimising free energy is considered a condition for life. One can equate 'survival' with 'life', since one supposes the other under the FEP; 'if I survived, it means that I maintained my structural integrity in face of environmental perturbations'; 'maintaining my structural integrity is what qualifies me as living'.

Now, it might be said that metamorphic organisms, despite not keeping their structural integrity, should be considered as living organisms. This was noted by Kirchhoff et al. 2018 and Clark 2017. From the point of view of the FEP, when considering such metamorphic organisms, it may be said that it is the lifecycle that corresponds to the thing whose integrity is maintained over time (e.g., over evolutionary time), not the specific form that the system takes at one stage of its development (e.g., the adult form of a frog). This casts the FEP within the realm of process ontology (vs. substance ontology). Similarly, while it may be argued that life is a state instantiated by an organism at time 't' whereas survival is a process with duration (e.g., the endurance of life from t to $t+1$), under the FEP (from the perspective of process ontology), it is unclear whether these two notions—life and survival—really have a different referent. It might be argued that both life and survival refer to an enduring process that advocates of process ontology would call 'organism' (cf. Dupré 2020). While there is certainly a fuller discussion to be had concerning the correct unit of analysis for free energy minimising organisms and the meaning of life and survival under that theory, a *critical* discussion of these issues is beyond the scope of this paper.

The first type of claim on the relationship between life and free energy minimisation is a strong type according to which minimising free energy is a sufficient condition for life. This claim has been called the overly generous claim (Kirchhoff and Froese 2017). Such a claim is attractive since it suggests that knowing what is involved in minimising free energy (e.g., possessing a Markov blanket) will inform us about what it takes to be alive. Such a strong claim would allow us to generalise the scope of the FEP to the full range of possible beings, and in so doing, it would allow us to predict which of those will pass the bar for qualifying as 'living';

it would allow one to identify 'what it takes' to be alive from the point of view of the FEP.

The second type of claim is a weak type according to which if a system is currently alive, it means that it minimised its free energy. Such a type of claim does not assume that the FEP is designed to set the bar for the sufficient conditions for life or meant to predict what things may or may not be alive. Rather, it limits the scope of application of the principle to beings that we think are alive, now, and enables us to know the necessary conditions under which those beings can be living—i.e., can actively resist the loss of structural integrity; 'what took them there'.

In the primary literature on the FEP, we can often read passages that may be interpreted as making strong claims, such as: "minimisation of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" (Friston and Stephan 2007, p. 26), and "systems that do not minimize free energy cannot exist" (Friston 2013, p.2). And so, people have reacted saying things of the sort "the right direction of explanation must go from minimizing free energy to survival. Yet insofar as FEP implies a causal story about that direction of explanation, it appears to be wrong. On the one hand, minimizing free energy cannot be sufficient for survival" (Klein 2018, p. 12). Here, Klein advocates the impossibility of a strong claim against the FEP. In the secondary literature, claims such as the aforementioned ones found in the primary literature have led some people to claim that the goal of the FEP is to discover the necessary characteristics of living systems, and that the free energy minimisation is an 'imperative' of life (Van Ess 2020). Here, one might argue that the terms 'imperative' and 'necessary' are correctly employed in the weak sense—i.e., in the sense of 'if life, then free energy minimisation has occurred'—but not a sufficient one, in the sense of 'if free energy minimisation occurs, then life follows'. But had the relation between life and free energy minimisation been correctly interpreted as merely necessary, some of the problems that van Es' claim is meant to motivate would simply not apply. Indeed, although it is hard to find direct evidence of what I called the entailment problem in the literature. That problem often transpires through some of the challenges that motivate philosophers to write on the FEP.

Take for instance the problem of scope, which is considered a serious problem among others by van Es. The scope problem refers to the danger of being over generous with applications of the FEP, out of fear of being overly generous with what we count as living (or as having a mind). Obviously, this is only a problem for someone who thinks that the FEP is meant to provide sufficient conditions for life (or mind). For instance, referring to a passage of Karl Friston's seminal paper 'Life as we know it' (2013), Kirchhoff and Froese (2017) say that:

> Strictly speaking, what Friston says here is that for any system to exist it must work to minimize free energy. This commits Friston to one of the following three implications. First, if free energy minimization is sufficient for mentality,

> then every system has a mind, even if not all systems are alive. Second, if free energy minimization is enough for life and mind, then all systems that exist are both alive and mental. Finally, biological systems, like all other existing systems, need to work to minimize free energy. The last option states that free energy minimization is not a property of only living systems, and as such sets up one of the two following implications. Either (option one) the FEP places mentality in a class of systems that includes but is not limited to living systems, and therefore veers toward some form of panpsychism. Or (option two) the FEP equates life-mind continuity with a view that sees life and mind nearly everywhere. […]. Our point is: given that the core concepts of non-cognitivist FEP—approximate Bayesian inference, ergodicity, Markov blankets and so on—can be applied to living and cognitive systems, on the one hand, and seemingly non-living and non-cognitive systems, on the other, there is a clear danger of these concepts being over-broad in their application, resulting in either seeing life and mind nearly everywhere or in the FEP lacking explanatory power when having to address the nature of life and mind and their relation to one another" (Kirchhoff and Froese 2017, pp. 10–11).

There is no such danger associated with the FEP for the simple reason that it is not because a system minimises its free energy (and has a Markov Blanket) that that system is alive. Again, free energy minimisation is not a sufficient condition for life (or mind). It seems to me that the problem of scope would only worry those who believe that the FEP makes a strong, sufficiency claim about the relation between life (or mind) and free energy minimisation.

Other standard manifestations of the entailment problem take the form of a critique of the 'testability' and 'tautology' of the FEP, which would be worries for the strong claimers, and for people who are generally worried about the explanatory power of the FEP, as mentioned by Kirchhoff and Froese. I do not have the space to elaborate on this here, plus this has already been done (Colombo and Wright 2018). Instead, in this paper, I simply dissolve what I called the entailment problem by providing a numerical example of free energy minimisation in a hypothetical organism (for a complete example, see Tschantz et al. 2020). I conclude with some brief epistemological remarks that may be of interest for those who worry about the explanatory power of the FEP.

The proposed numerical example will clearly demonstrate why minimising free energy can generate both Bayesian adaptive and Bayesian maladaptive behaviour, leading to survival, or death, accordingly. The proposed numerical example demonstrates that minimising free energy is not sufficient for life—the strong claim. The proposed numerical example does not demonstrate the necessity claim; the idea that under the right conditions, remaining alive means that free energy was minimised—the weak claim. However, the weak claim should become apparent through the reading of the numerical example, which will show that under the right conditions, minimising free energy should allow the maintenance of structural integrity. Hopefully, this numerical example will appease those who want to raise worries, implicitly or explicitly, about the—non-existent—FEP strong claim, or about the apparently less interesting weak claim.

Minimising free energy: for better or worse

Some conceptual distinctions between Bayes and the free energy principle

Bayesian approaches to animal behavior propose that one can model organisms as representing their relation to environmental states using priors and a likelihood (McNamara et al. 2006). Let's call those representations Bayesian 'beliefs'. On the basis of those beliefs, organisms generate adaptive behaviour. Bayesian beliefs represent (i) the probability of environmental states, prior to observing an environmental signal (a.k.a. prior); and (ii) the relation between environmental states and observed environmental signals (a.k.a. likelihood). Bayes theorem, from which terms such as prior and likelihood come from, is typically expressed as an evidentiary relationship between some prior hypotheses (P(H)) and the observation at hand, or data (E): $P(H|E) = [P(E|H)P(H)]/P(E)$.

The free energy principle is a Bayesian formulation of the manner in which organisms infer the posterior probability of their prior beliefs after having observed an environmental signal with a given likelihood, and in so doing infer some hidden, or unobserved variables. What stands for the 'H' are the unobserved variables whose prior probability P(H) forms the hypothesis, and what stands for the 'E' are the sensory signals organisms receive (the data). Hence, it is often said that under the free energy principle, organisms are viewed as embodying an 'hypothesis', a 'belief' or a 'best guess' about the cause of their sensations, or sensory signals they receive (Allen and Friston 2016; Bruineberg and Rietveld 2014; Friston 2011).

Under the FEP, the evidentiary relation explains the manner in which organisms selfevidence (Hohwy 2016), where the 'self' means evidencing beliefs about oneself in the world. Because beliefs are embodied by the organism, and are thus the organism's own states, the uncertainty in the likelihood and the prior can be viewed as representing the uncertainty inherent to the biological apparatus (e.g., noise in the signal transmission across the nervous system), instead of the uncertainty of the world (e.g., fluctuations in states of the world generating the signals), as would be the case under typical Bayesian models. Under the FEP, uncertainty should thus be read as reporting a Bayesian 'credence score' over the organism's own beliefs, as it reports the probability of a state or hypothesis relative to other possible hypotheses. In the case of the likelihood, the credence score is over sensory beliefs relative to states (e.g., 'is this more probably warm or more probably hot?'). In the case of the prior, the credence score is over the hypotheses the organism entertains prior to sensing the water temperature (e.g., 'am I probably in the ocean or in my bath?'). Of course, these beliefs, hypotheses, or best guesses are implicit and subpersonal, as they are meant to be realised by the organism's (neuro)physiology. This begs another important question: are priors subjective or objective under the FEP?

Priors can be of two kinds: (i) objective, or (ii) subjective. Objective priors are typically based on frequencies (e.g., priors that reports distributions based on

empirical data). When the frequencies are unknown, an equiprobable (flat) prior should be favoured. Objective priors thus conform to some rational constraints beyond Bayesian rationality. Subjective priors, in turn, refer to the psychological dispositions of the system of interest, or to the person specifying the system of interest (e.g., priors that reports propositional attitudes). Subjective priors do not need to conform to constraints of rationality. The simple answer to the question of whether priors are objective or subjective under the FEP is that they are subjective. As we said above, they track the confidence over a system's beliefs. We could thus carry on with that in mind.

However, there is an interesting detail on that question that may worth mentioning. Under the FEP, priors do not conform to a rationality beyond the rationality of the inference per se, but nor are they rationally unconstrained. There is a rationality beyond Bayesian rationality that comes from the way variational Bayes is realised by the system (Hohwy 2020). As we will see in detail later on, the inference of the posterior distribution requires finding an approximation to that posterior (denoted as 'Q' later on), which then becomes the prior used in the next cycle of inference. That approximate posterior determines what is embodied by the organism. The update having led to the subjective posterior at time $t+1$ operates by finding the subjective posterior that would *best* approximate the true subjective posterior distribution. The meaning of 'best' just is being close to 0 free energy. That true subjective posterior is that which one would find with exact Bayesian inference—more on this later, and crucially never exist. Hence it is sometimes said that it forms only a reference point to perform the inference (Ramstead et al. 2019).

The rational constraint over the priors is the fact that the approximate subjective posterior 'Q' (or future prior) will not only be Bayesian, but also will always be the 'best guess' relative to what the true posterior *ought to be*. In short, under the FEP, even though priors refer to psychological states of the system, updates of the system make those priors an approximation of what they 'should' have been, had the prior been updated with exact Bayes. Thus, it might be said that priors under the FEP cut across the objective / subjective dichotomy. They are subjective while satisfying a rational constraint mandated by the existence of the system per se.

The numerical example

The numerical examples below operate under the following scenario (see Fig. 1). Consider an organism that infers whether an external event A or B took place. For the organism, A and B are part of the class R and form the representations by the organism of the external events A or B. A and B are inferred when receiving a chemical signal part of the class S, which can be alpha or beta.

We assume that before observing any signal, the probability of A is p, and the probability of B is 1-p. Given the environment in which the organism finds itself, the probability of observing a signal alpha under A is m, and the probability of observing beta under A is 1- m. We assume that p is equal to 0.8, and that m is equal to 0.7. The opposite applies to B. We stipulate for the sake of the numerical example that representing A when receiving alpha, or B when receiving beta leads

Generative model (a.k.a. joint probability) $= P(R,S) = P(R)P(S \mid R)$

$$\textbf{Prior} = P(R) = \begin{bmatrix} .8 \\ .2 \end{bmatrix} \begin{matrix} A \\ B \end{matrix}$$

$$\textbf{Likelihood} = P(S \mid R) = \begin{matrix} \\ \alpha \\ \beta \end{matrix} \begin{matrix} A & B \\ \begin{bmatrix} .7 & .3 \\ .3 & .7 \end{bmatrix} \end{matrix}$$

$$\textbf{Bayesian organism} : P(A \mid \alpha) = \frac{P(A)P(\alpha \mid A)}{P(A)P(\alpha \mid A) + P(B)P(\alpha \mid B)}$$

$$\textbf{Variational Bayesian organism} : Q(A) = \arg\min_{Q} F \Rightarrow Q(A) \approx P(A \mid \alpha)$$

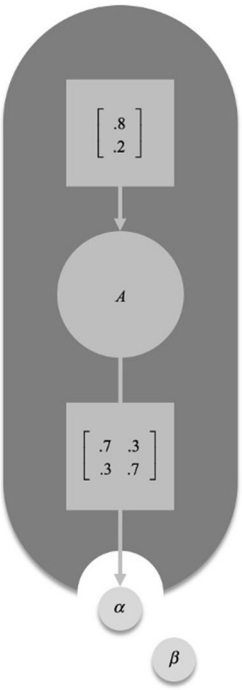

Fig. 1 Right panel.: Visual representation of an organism inferring its prior beliefs about the cause of the observation that it makes. The prior beliefs are assumed to complement the external cause of the observation. Here, the organism observes the outcome 'alpha', and on the basis of its prior and likelihood (i.e., sensory beliefs) finds the posterior value of its beliefs. Given that behaviour is formally equivalent to inference in our simple organism, inferring 'A' as the right beliefs about the most probable cause of observation means biophysically representing 'A'. There is no action involved in our example. The likelihood and the prior are assumed to 'map', heuristically, onto the physiology of our organisms—the prior being some sort of storage of knowledge, and the likelihood being the sensory belief. In more biologically realistic descriptions of behaviour, which require a discussion of active inference, behaviour is the result of a different inference process—that of an action policy (under discrete models). This involves more priors, namely, about the transition between hidden states and often about preferred sensory outcomes. Action then is distinguished from inferring hidden states. It is about inferring another hidden variable, which is the policy. Left panel: The first line represents the organism, formally, as a joint distribution obtained by multiplying the prior and the likelihood (which is biophysically implemented). This joint distribution can also be viewed as a 'generative model', or model of the manner in which sensations are caused by external states. Inferring the posterior probability, based on that joint distribution or generative model, allows the organism to respond adaptively and to generate for itself the right sensation. Indeed, one must distinguish the sensory input (e.g., alpha or beta) from the generated sensation by the organism. The second and third lines represent the prior and the likelihood, formally. The fourth line represents the possible Bayesian algorithm that could be used. The fifth line presents variational free energy minimisation that selects the approximate posterior density (Q(A))

to survival, and that the opposite leads to death. Because inference is biophysically realised, representing, or inferring A or B could also be interpreted as producing a metabolic response (not necessarily an action) to alpha or beta. Heuristically, the reader can assume that we simply stipulate that inferring B when sensing alpha or A

when sensing beta is a maladaptive metabolic response that prevents from maintaining structural integrity. The prior probability P(R) and the likelihood P(S|R) can be visualised as follows:

$$
\begin{aligned}
&P(A) = p = .8\\
&P(B) = 1 - p = .2
\end{aligned}
$$

$$
\text{Prior: } P(R) = \begin{bmatrix} .8 \\ .2 \end{bmatrix}
$$

$$
\begin{aligned}
&P(\alpha|A) = m = .7\\
&P(\beta|A) = 1 - m = .3\\
&P(\alpha|B) = 1 - m = .3\\
&P(\beta|B) = m = .7
\end{aligned} \tag{1}
$$

$$
\text{Likelihood: } P(S|R) = \begin{bmatrix} .7 & .3 \\ .3 & .7 \end{bmatrix}
$$

Assuming that the internal computation that our organism performs conforms to Bayes theorem (McNamara et al. 2006; Okasha 2013), computing the posterior probability of A or B relative to the environmental signal amounts to representing the most likely state. Let's infer the posterior probability of A after observing, say, alpha. To do this, we would apply Bayes theorem as follows:

$$
\begin{aligned}
&P(A) = .8\\
&P(\alpha|A) = .7\\
&P(B) = .2\\
&P(\alpha|B) = .3\\
&\underbrace{P(A|\alpha) = \frac{P(\alpha, A)}{P(\alpha)}}_{\text{Bayes rule}}\\
&\underbrace{P(\alpha, A) = P(A)P(\alpha|A) = .8 * .7 = .56}_{\text{Joint probability of } A \text{ and } \alpha}\\
&\underbrace{P(\alpha) = P(A)P(\alpha|A) + P(B)P(\alpha|B) = .56 + .06 = .62}_{\text{Marginal distribution (a.k.a. model evidence)}}\\
&P(A|\alpha) = \frac{P(\alpha, A)}{P(\alpha)} = \frac{.56}{.62} = .9032
\end{aligned} \tag{2}
$$

Equation (2) takes the prior probability of A, which is 0.8, and multiplies it by the likelihood of A under signal alpha, which is 0.7, in order to get the joint probability of A and alpha, which is 0.56. In order to find the posterior probability, one must

divide this joint probability by the marginal distribution, which is simply the sum of the joint probability for A and B under signal alpha, respectively; or alternatively, the prior for B times the likelihood for B under alpha, plus the prior for A times the likelihood for A under alpha. Exact Bayesian inference yields a posterior probability of 0.9032 for state A after having observed the signal alpha (and a posterior of 0.0968 for B, since the posterior distribution must sum to 1). This means that after seeing alpha, an exact Bayesian organism would have represented A with ~ 90% confidence, and thus would have survived.

With exact Bayesian inference, one uses the marginal distribution to find the posterior probability. This assumes that the organism could sum over the probability of outcomes under both A and B. However, it is unclear whether living systems have sufficient computational power to accomplish that (Bogacz 2017; Friston 2009). For instance, following our numerical example, the signal alpha might have been caused by environmental states A,B,C,…, each of which would have an analogue internal state A,B,C represented by the organism. Thus, the likelihood modelled by the organism might look like this:

$$\underbrace{P(S|R)}_{\text{Likelihood}} = \begin{bmatrix} P(\alpha|A) & P(\alpha|B) & P(\alpha|C) & \ldots \\ P(\beta|A) & P(\beta|B) & P(\beta|C) & \ldots \end{bmatrix} \tag{3}$$

Under exact Bayesian inference, all the probabilities in Eq. 3, for all states under the observation of interest (e.g., alpha) should be summed over. Doing this will often be computationally intractable, as the organism will entertain multiple different causal representations (e.g., A,B,C…) for the same observation (e.g., a red sensation that might have been caused by a red 'shoe', red 'car', red 'traffic light', red …). This problem underwrites what is referred to in the literature on the free energy principle and predictive processing as the black box problem (Clark 2013), the solipsism problem (Hohwy 2016), or the seclusion problem (Wiese and Metzinger 2017).

In order to bypass this problem, the FEP models the inference process (e.g., of A or B) performed by organisms as approximate Bayesian inference. Approximate Bayesian inference bypasses the direct evaluation of the likelihood and the marginal distribution when inferring the posterior probability. Note that in biology, similar methods became popular through work in population genetics on the genealogy of DNA sequences (Sunnåker et al. 2013; Tavaré et al. 1997). The central claim of the FEP is that changes leading to behavioral and (neuro)physiological responses in living systems conform to a form of approximate Bayesian inference known as variational Bayes (Beal 2003; Friston 2005, 2013; Parr and Friston 2018).

Now, building on the numerical example above, the following numerical example shows that one can infer the posterior probability for A by minimising free energy; and with the same inference process and the same likelihood, one can find a posterior that gives high confidence to B. Given that representing A when observing alpha leads to survival, and representing B when observing alpha leads to death, the following numerical example will demonstrate that minimising free energy is not a sufficient condition for life, as it can lead to the exact opposite—death.

Note that the scope of the following numerical example is deliberately limited. The goal is to demonstrate that minimising free energy can lead to maladaptive inference when performed with the wrong priors, all things being kept fixed. If the priors are allowed to update, the inference should lead to adaptive behavior. This is an important point to which we will come back below. Adaptivity is guaranteed by the extent to which the priors match the environmental constraints, more than by the nature of the machinery employed to perform the inference (e.g., free energy minimisation or exact Bayes). That being said, the machinery that allows the inference will play an important role in allowing priors to match environmental constraints. The following example of free energy minimisation is provided in the sole purpose of supporting our response to the entailment problem. The goal is to give a formal intuition as to why minimising free energy is not sufficient for life understood as the preservation of structural integrity. By no mean should the following numerical example be viewed as an exemplar of the manner in which free energy minimisation operates, mathematically. The following numerical example simply illustrates the concepts engaged in this paper and does not provide a complete understanding of the mathematical apparatus of the FEP. Technical readers should refer to Buckley et al. (2017) and Bogacz (2017) or Smith et al. (2021).

Free energy 'F' is defined as follows:

$$F = -\sum_{R} Q(R)\left[\ln \frac{P(R,\alpha)}{Q(R)}\right] \tag{4}$$

Equation (4) says that free energy on the left side of the equation is equal to the (negative) sum of the log ratio of an approximation to the posterior for A and B (Q(R)) and the joint probability of those states and signal 'alpha' (P(R, alpha)), multiplied by the approximate posterior (Q(R)). Minimising free energy, from the perspective of Eq. 4, just means finding the approximate posterior Q(R) that will yield the F that is the closest to 0 on the left side of the equation. Q(R) corresponds to the proposal, recognition, or approximate posterior density sometimes referred to in the literature on the FEP. It is that Q(R) that is embodied by the organisms—not to confuse with the P(R,alpha), which would be the joint distribution, or generative model (Ramstead et al. 2019a, b).

Above, using exact Bayesian inference, we had to divide the joint probability of A and alpha by the marginal distribution. Recall that here, we want to remain agnostic concerning the marginal distributions to which we do not have access. We can find the posterior under such constraints by asking 'what approximate posterior Q(R) gives me the least F'? The answer to that question is the approximate posterior Q(R) that will be the closest to the true posterior.

We know from exact Bayes that the true posterior probability of A given alpha (P(A | alpha)) is 0.9032, meaning that after observing alpha, our exact Bayesian organism represented state A with ~ 90% confidence. Now, let's assume that our organism operates under variational Bayes, and that it indeed represented A with the same level of confidence. What would have been its free energy? This can be computed as follows:

$$
\begin{aligned}
F = & \quad -\sum_{R} Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right] \\
= & \quad \left(-\left(\underbrace{.9032}_{Q(A)} * \ln\frac{\overbrace{.56}^{P(\alpha,A)}}{\underbrace{.9032}_{Q(A)}}\right)\right) + \left(-\left(\underbrace{.9068}_{Q(B)} * \ln\frac{\overbrace{.06}^{P(\alpha,B)}}{\underbrace{.9068}_{Q(B)}}\right)\right) \\
= & \quad .4780
\end{aligned}
\tag{5}
$$

Equation 5 tells us that the free energy of an organism with an approximate posterior equal to the true posterior would be 0.4780; or put another way, minimising free energy down to 0.4780 means representing A with a level of confidence of ~ 90%. Now let's imagine an organism that would have inferred P(A | alpha) with a probability of 0.0968, which we know is far from the true posterior:

$$
\begin{aligned}
F = & \quad -\sum_{R} Q(R)\left[\ln\frac{P(\alpha,R)}{Q(R)}\right] \\
= & \quad \left(-\left(.0968 * \ln\frac{.56}{.0968}\right)\right) + \left(-\left(.0932 * \ln\frac{.06}{.0932}\right)\right) \\
= & \quad 2.2792
\end{aligned}
\tag{6}
$$

Equation 6 tells us that an organism that would have represented B with ~ 90% confidence after seeing alpha would have had a free energy of 2.2792, which is higher than 0.4780. Based on the current scenario (i.e., B when receiving alpha leading to death), the organism with the higher free energy would have died. Hence, one might be tempted to agree with the claim that minimising free energy is sufficient, if not necessary for survival. Indeed, when comparing eqs. 5 and 6, minimising free energy—i.e., finding the approximate posterior that yields the free energy closest to 0—guarantees survival, whereby the opposite guaranteed death.

However, minimising free energy leads to survival only under the right conditions, that is, if the organism has the right prior beliefs, and the right joint probability, accordingly. Let's imagine the same scenario, with the same likelihood and success conditions, but with inverted prior beliefs. This is conceivable, for instance, if an organism inherits maladaptive prior beliefs (Richerson 2018). Let's imagine that our organism has inherited a maladaptive, inverted prior:

$$
\begin{aligned}
&P(A) = .2\\
&P(B) = .8\\
&P(\alpha|A) = .7\\
&P(\alpha|B) = .3\\
\\
&P(\alpha, A) = .2 * .7 = .14\\
&P(\alpha, B) = .8 * .3 = .24\\
\\
&\text{Posterior of A} = .3684\\
&\text{Posterior of B} = .6316
\end{aligned}
\tag{7}
$$

Equation 7 Simply inverts the prior probability we started with in Eq. (1). and shows the consequence for exact Bayesian inference. With the same likelihood, but an inverted prior, an exact Bayesian organism would have represented state B with ~ 0.63% confidence after seeing alpha; and thus, would have died. As you might suspect it, the same applies to a free energy minimising organism:

$$
\begin{aligned}
F &= -\sum_R Q(R)\left[\ln \frac{p(\alpha, R)}{Q(R)}\right] \\
&= \Big(-\Big(.3684 * \ln \frac{.14}{.3684}\Big)\Big) + \Big(-\Big(.6316 * \ln \frac{.24}{.6316}\Big)\Big) \\
&= \underbrace{.9676}_{\text{F when representing B after sensing } \alpha}
\end{aligned}
\quad \leq \quad
\begin{aligned}
F &= -\sum_R Q(R)\left[\ln \frac{p(\alpha, R)}{Q(R)}\right] \\
&= \Big(-\Big(.6316 * \ln \frac{.14}{.6316}\Big)\Big) + \Big(-\Big(.3684 * \ln \frac{.24}{.3684}\Big)\Big) \\
&= \underbrace{1.094}_{\text{F when representing } A \text{ after sensing } \alpha}
\end{aligned}
\tag{8}
$$

Equation 8 tells us that when observing alpha, representing B with ~ 0.63% confidence yields a free energy of 0.9676, which is closer to 0 than 1.1094. This means that an organism minimising its free energy would have represented B instead of A when observing alpha. In the current scenario, this is fatal. Hence minimising free energy per se does not entail life; not under the wrong prior. In fact, it can perfectly well entail the exact opposite. And so, it should be clear that claims according to which free energy minimisation provides the sufficient conditions for life should not be interpreted as such. There is no logical consequence that goes from minimising free energy to life understood as maintaining one's structural integrity.

Free energy on a wing and a prior?

Although free energy minimisation is not sufficient for life, the numerical example above suggests that there might be an entailment relation that goes the other way around: if you are alive, it might very well because you did something like minimising free energy. That entailment relation is that which corresponds to the weak version of the life-free-energy entailment relation. Indeed, in our numerical example (Eqs. 5 and 6), minimising free energy led to survival under the right (prior) conditions; and it seems fair to assume that from a Bayesian point of view, minimising free energy (or performing a similar form of approximate inference) is what

organisms do. This makes the FEP an interesting epistemic principle for researchers interested in development. In a reverse engineering fashion, if we observe a free energy minimising organism that is still living at the time that we observe it, we can trust that it has good enough priors to remain alive; and if we observe that that organism behaves maladaptively, we have good reasons to doubt the viability of its current priors. The goodness of priors, of course, rests on the extent to which priors match the sort of challenges the organism is currently exposed to (e.g., if you represent 'B' when sensing 'alpha', you die, and so a good prior is a prior that makes you represent A more often than not—has higher credence on A). The consequence of this is that one can bring the sufficiency claim back into the game if one assumes that the organism is endowed with adaptive prior beliefs.

One could rightfully say that minimising free energy when equipped with the right prior beliefs is sufficient for life; i.e., that it is all you need to be qualified as living, or as maintaining your structural integrity under the free energy principle. The point here is that the choice of prior, whether under Bayesian or approximate Bayesian regimes is the real concern, since the entailment relation between life and the FEP entirely depends on the adaptivity of those priors. Assuming that priors are genetically inherited, the entailment relation between the FEP and life will be predicated on evolutionary processes. Interestingly, some have argued that the adaptivity of priors can also be guaranteed by free energy minimisation operating at the population level, as a form of natural selection (Badcock et al., 2019; Constant, Ramstead, et al., 2018; Friston 2010, 2013; Friston and Stephan 2007; Hesp et al. 2019; Ramstead et al. 2017; Sella and Hirsh 2005). To make sense of this, simply imagine that instead of modelling an organism with states A and B minimising free energy, we are modelling a population with different genotypic states AA Aa aa, each having a prior probability, and each being more or less likely under observable environmental patches. Minimising free energy at the population level would allow natural selection to converge on the Bayesian gene pool distribution, that is, the approximate posterior distribution for genotypes that is the closest to the true posterior distribution under reproductive observations. This means that inherited genotypic priors sampled from the approximate posterior distribution at the genotypic level should be well tuned to the environmental pressures that have caused the reproductive success (i.e., observations). By extension, individuals having received the most probable genotype will have genotypic priors that provide the right prior conditions for successful behaviour (e.g., representing A when observing alpha).

However, even such a multiscale free energy minimisation rationale does not guarantee that organisms with the right inherited priors won't undergo somatic mutations, or simply neural lesions that would change the distribution of inherited priors, therefore biasing free energy minimisation over development toward faulty inference, death and the inability to maintain structural integrity.

Future direction: free energy minimisation as a historical scientific principle?

The dissolution of the entailment problem puts us in a good position to move on to another related difficulty in the philosophical literature on the FEP, which is, this time, of an exegetic kind. If minimising free energy is not sufficient for life or survival, how should we interpret statements such as "the minimisation of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" (Friston and Stephan 2007, p.428)?. I conclude with an epistemological remark on the meaning of that statement.

The FEP on its own is a principle, namely, a foundation for reasoning about things (e.g., living things). In this paper, we approached the FEP as such. However, the FEP can also be read more broadly as a research program that uses FEP reasoning patterns to generate scientific hypotheses. This involves implementing FEP reasoning into a theory called active inference, which is routinely used to study various cognitive functions (for a review see Da Costa et al. 2020). As a research program, the FEP can be used to generate statements that are normative in the strong sense. Such statements can be tested using scientific standards for hypothesis testing (Smith et al. 2020, 2021).

As a reasoning pattern, the FEP can be used to generate postdictive statements (cf. Friston et al. 2017).[1] Accordingly, FEP reasoning might be interpreted as a principle akin to those found in postdictive sciences (a.k.a. historical sciences) like geology, palaeontology, archaeology, or any science that deals with irreproducible causes (Cleland 2002). Postdictive scientific statements are concerned with what 'must have been the case', instead of 'what will be' the case. A statement such as "the minimisation of free-energy may be a necessary, if not sufficient, characteristic of evolutionary successful systems" is probably such a postdictive statement. That statement should be interpreted as claiming that free energy minimisation must have occurred if a system is evolutionarily successful—not the other way around. Nonetheless, this is an interesting statement because if free energy has occurred, the system in question can be modelled as if it possesses the features allowing for free energy minimisation (e.g., a Markov Blanket). One can then start inquiring about whether those features help us understand the sort of dynamics implemented by the (neuro)physiology of the system, in a predictive fashion (e.g., with the FEP as a research program). Hence, it is sometime said that the FEP, as a foundation for reasoning, is a 'guide to discovery' (Ramstead et al. 2017).

According to Cleland (2002), historical scientific methodology enables scientists to generate historical hypotheses about the best causal explanation for some observations, based on the accumulation of evidence about the causal structure that might

[1] It is important to note that the FEP includes processes other than free energy minimisation. It also includes expected free energy minimisation (and generalised free energy minimisation, (Parr and Friston 2019)). While minimising free energy endows the organism with postdictive inference, minimising expected free energy endows the organism with predictive inference. This is due to the simple reason that the outcomes and states involved in the inference process under expected free energy minimisation are in the future, not the present. Effectively, this means that inferring one's beliefs about states of the world means inferring what will most likely be seen under those beliefs, and under a given sequence of action to be engaged (i.e., action policy).

have led to those observations (e.g., evidencing the asteroid-impact hypothesis of dinosaurs' extinction using fossil records of asteroid's impact). In historical sciences, an 'investigator' starts by observing some puzzling traces, or the effects of a cause in the distant or proximal past. The investigator then postulates some hypotheses about the cause of the observed effects. Testing a historical hypothesis then just means accumulating more traces to evidence one of the competing historical hypotheses. These new traces are 'smoking guns', which are meant to shift the 'balance of probability' towards one of the competing hypotheses. A historical hypothesis is defined by the pattern whereby it is evidenced and by its ability to account for those smoking guns with a unifying and compelling causal story.

FEP reasoning yields historical hypotheses because it operates a historical evidentiary pattern and provides a compelling unifying causal story. It operates a curious evidentiary pattern, though, because it assumes that both the investigator and the thing under investigation conform to that evidentiary pattern. That pattern is free energy minimisation, per se. For instance, for the organism in our numerical example, the hypotheses were A or B. The smoking guns were the sensory observations 'alpha' or 'beta'. The (self)evidencing activity whereby the organism 'tested' those hypotheses was biophysically realised variational Bayes using the sensory observation to evidence the hypotheses about itself (e.g., A; B). Then, as a person who used the free energy principle in the numerical example above, the puzzling trace for which I was seeking a causal explanation was the survival of the organism. That was my observation. The causal story or hypothesis for that observation under the conditions we imposed to our simulated organism was the free energy principle, the inference over which led me to write the paper you are reading at the moment. That paper functioned as sensory evidence for my hypothesis (e.g., when writing down the number and seeing they were adding up). And that paper is the observation that you are using to evidence your hypotheses concerning the claim I set at the start of the paper, namely, that free energy minimisation is not sufficient for life. Fidel to the unifying grip of hypotheses in historical sciences, the free energy principle is meant to account for all of that—you, me and the organism under study, in a unifying fashion.

Acknowledgements I want to thank Paul Badcock, Paul Griffiths, Mark Colyvan, Christopher Whyte, Pierrick Bourrat, Joshua Christie, Christopher Lean, Peter Takacs, Carl Brusse, Stefan Gawronski, and Walter Veit for helpful comments on earlier versions of this paper.

Funding Work on this article was supported by the Australian Laureate Fellowship project A Philosophy of Medicine for the 21st Century (Ref: FL170100160) and by a Social Sciences and Humanities Research Council doctoral fellowship (Ref: 752–2019-0065).

References

Allen, M., Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. Synthese, 1–24.

Badcock PB, Davey CG, Whittle S, Allen NB, Friston KJ (2017) The depressed brain: an evolutionary systems theory. Trends Cognit Sci 21(3):182–194

Beal MJ (2003) Variational algorithms for approximate bayesian inference. University of London, London

Bogacz R (2017) A tutorial on the free-energy framework for modelling perception and learning. J Math Psychol 76(Pt B):198–211
Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. J Math Psychol, 81(Supplement C), 55–79.
Bruineberg J, Rietveld E (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. Front Human Neurosci 8:599
Clark A (2013) Whatever next? predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36(03):181–204
Clark A (2017) How to knit your own Markov blanket: resisting the second law with metamorphic minds. In Philosophy and predictive processing: 3 (eds Metzinger T, Wiese W). Frankfurt am Main, Germany: MIND Group.
Cleland CE (2002) Methodological and epistemic differences between historical science and experimental science*. Phil of Sci 69(3):447–451
Colombo M, Wright C (2018) First principles in the life sciences: the free-energy principle, organicism, and mechanism. Synthese. https://doi.org/10.1007/s11229-018-01932-w
Constant A., Ramstead MJD, Veissière SPL, Campbell JO, Friston KJ (2018). A variational approach to niche construction. J R Soc Interface R Soc, 15(141). https://doi.org/10.1098/rsif.2017.0685
Da Costa L, Parr T, Sajid N, Veselic S, Neacsu V, Friston K (2020). Active inference on discrete state-spaces: a synthesis. In arXiv [q-bio.NC]. arXiv. http://arxiv.org/abs/2001.07203
Dupré J (2020) Life as process. Epistemol Phil Sci 57(2):96–113
Friston KJ (2005) A theory of cortical responses. Phil Trans R Soc London Series B Biol Sci 360(1456):815–836
Friston KJ (2009) The free-energy principle: a rough guide to the brain? Trends Cognit Sci 13(7):293–301
Friston KJ (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11(2):127–138
Friston KJ (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). Imprint Academic.
Friston KJ (2013) Life as we know it. J R Soc Interface R Socy 10(86):20130475
Friston KJ, Parr T, de Vries B (2017) The graphical brain: Belief propagation and active inference. Netw Neurosci 1(4):381–414
Friston KJ, Stephan KE (2007) Free-energy and the brain. Synthese 159(3):417–458
Friston KJ, Thornton C, Clark A (2012) Free-energy minimization and the dark-room problem. Front Psychol 3:130
Hesp C, Ramstead MJD, Constant A., Badcock P (2019). A multi-scale view of the emergent complexity of life: a free-energy proposal. Evolution & Development. https://link.springer.com/chapter/https://doi.org/10.1007/978-3-030-00075-2_7
Hohwy J (2016) The self-evidencing brain. Noûs 50(2):259–285
Hohwy J (2020) Self-supervision, normativity and the free energy principle. Synthese. https://doi.org/10.1007/s11229-020-02622-2
Kirchhoff M (2015) Species of realization and the free energy principle. Australas J Philos 93(4):706–723
Kirchhoff M, Froese T (2017) Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy*, *19*(4): 169.
Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. Journal of the Royal Society, Interface / the Royal Society, *15*(138) https://doi.org/10.1098/rsif.2017.0792
Klein C (2018) What do predictive coders want? Synthese 195(6):2541–2557
McNamara JM, Green RF, Olsson O (2006) Bayes' theorem and its applications in animal behaviour. Oikos 112(2):243–251
Okasha S (2013) The evolution of bayesian updating. Philos Sci 80(5):745–757
Parr T, Friston KJ (2018) The anatomy of inference: generative models and brain structure. Front Comput Neurosci 12:90
Parr T, Friston KJ (2019) Generalised free energy and active inference. Biol Cybern. https://doi.org/10.1007/s00422-019-00805-w
Ramstead MJD, Badcock PB, Friston KJ (2017) Answering Schrödinger's question: a free-energy formulation. Phys Life Rev 24:1–16
Ramstead MJD, Kirchhoff MD, Friston KJ (2019). A tale of two densities: active inference is enactive inference. Adapt Behav, 1059712319862774.

Ramstead MJD, Kirchhoff MD, Constant A, Friston KJ (2019) Multiscale integration: beyond internalism and externalism. Synthese. https://doi.org/10.1007/s11229-019-02115-x

Richerson PJ (2018) An integrated bayesian theory of phenotypic flexibility. Behav Proc. https://doi.org/10.1016/j.beproc.2018.02.002

Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. Proc Natl Acad Sci USA 102(27):9541–9546

Smith R, Friston K, Whyte C (2021). A step-by-step tutorial on active inference and its application to empirical Data. https://doi.org/10.31234/osf.io/b4jm6

Smith R, Kuplicki R, Teed A, Upshaw V, Khalsa SS (2020). Confirmatory evidence that healthy individuals can adaptively adjust prior expectations and interoceptive precision estimates. In Cold Spring Harbor Laboratory (p. 2020.08.31.275594). https://doi.org/10.1101/2020.08.31.275594

Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C (2013) Approximate bayesian computation. PLoS Comput Biol 9(1):e1002803

Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145(2):505–518

Tschantz A, Seth AK, Buckley CL, Komarova NL (2020) Learning action-oriented models through active inference. PLOS Comput Biol 16(4):e1007805

Van Es T (2020). Living models or life modelled? on the use of models in the free energy principle. Adapt Behav, 1059712320918678.

Wiese W, Metzinger T (2017) Vanilla PP for Philosophers: A Primer on Predictive Processing. https://philarchive.org/rec/WIEVPF?all_versions=1