Minds and Machines (2024) 34:29
https://doi.org/10.1007/s11023-024-09690-0

Measure for Measure: Operationalising Cognitive Realism

Majid D. Beni[1]

Received: 28 October 2022 / Accepted: 24 June 2024

Abstract
This paper develops a measure of realism from within the framework of cognitive structural realism (CSR). It argues that in the context of CSR, realism can be operationalised in terms of balance between accuracy and generality. More specifically, the paper draws on the free energy principle to characterise the measure of realism in terms of the balance between accuracy and generality.

Keywords Cognitive structural realism · Free energy principle · The strategy of model-based science · Realism · Operationalism

1 Introduction

The free energy principle (FEP), which is articulated by Karl Friston and colleagues, is at the centre of flourishing research streams in computational neuroscience and theoretical biology. There are vibrant debates over the right philosophical interpretation of FEP. These interpretations come in various flavours, ranging from outright instrumentalism (Colombo & Palacios, 2021; van Es & Hipolito, 2020) and moderate non-realism (Ramstead et al., 2020) to outright or moderate versions of realism (Beni, 2019b; Kirchhoff et al., 2022), with interesting remarks on how the modelling practice of FEP theorists may bear on the realist/antirealist interpretation of FEP (Andrews, 2021; Beni, 2021a; Friston et al., 2020). This paper goes beyond just defending realism about FEP and sets itself the more ambitious task of endorsing FEP as an operational measure that lies at the centre of a new take on scientific realism. To be more precise, the paper draws on FEP to characterise a measure for realism in terms of the balance between the generality and accuracy of scientific models.

Despite the novelty of the proposal of the paper—to use FEP to characterise a measure for scientific realism—the philosophical tradition that motivates our quest has been around for quite a while. The enterprise of this paper is inspired by the

✉ Majid D. Beni
mbeni@metu.edu.tr

[1] Department of Philosophy, Middle East Technical University, Ankara, Turkey

Published online: 05 July 2024

informational structural realism in the philosophy of science, in the works of Floridi (2008, 2009), Ladyman and Ross (2007). More specifically, the present paper is a derivation of a (cognitive) derivation of Floridi's realism—with the cognitive derivation being called Syntactical or more recently cognitive structural realism (CSR) (Beni, 2016, 2018, 2019a). Amongst other things, CSR has been claimed to underwrite a naturalist methodology of science (Beni & Pietarinen, 2021; Pietarinen & Beni, 2021). It also arguably accounts for social aspects of scientific practice on the basis of FEP's theory of (dyadic) alignment (Beni, 2021c). However, despite being introduced as a version of *realism*, CSR simply has not provided a clear statement of its realist tendency. The present paper aims to amend this shortcoming. CSR's account of scientific representations (which are supposed to be truthful, if CSR is a version of realism) builds on FEP. This paper relies on the theoretical resources of FEP once more to offer a measure for realism under CSR, which will be supplemented with a definition of realism in terms of the balance between the accuracy and generality of scientific models. To relate our measure of realism to the extant philosophy of science, the contribution of the paper will be presented as a revision of Levins's (1966) discussion of the strategy of model-based science.

The paper is structured as follows. Section 2 introduces the free energy principle and cognitive structural realism. Section 3 visits Levins's account of the relationship between generality and realism (articulated in terms of accuracy) and rehearses reasons for scepticism about identifying realism with accuracy. Section 4 shows how FEP provides an operational measure for finding the balance between accuracy and generality. Section 5 is the conclusion.

2 The Free Energy Principle and Cognitive Structural Realism

The free energy principle, as being developed by Karl Friston and colleagues (Friston, 2010; Friston et al., 2010; Ramstead et al., 2017), lies at the centre of a unifying theoretical framework for aspects of computational neuroscience, theoretical biology and physics. According to the second law of thermodynamics, the sum of entropies of all of the systems that attain thermodynamic equilibrium would increase. In order to survive, biological systems must defy the second law. They do so by minimising their variational free energy, which is defined based on 'self-information', 'surprisal', or simply 'surprise' (the relation between the statistical term surprisal and psychological surprise has been a point of contention, but we overlook the difference in this paper).

The mathematical articulation of FEP has been subject to various formulations over time. One neat way of modelling FEP (and the notion of relative entropy) consists in using Kullback–Leibler divergence (KL-divergence), assuming that minimising free energy corresponds to minimising the divergence between expected and actual (future) entropy of the self-organising system.

We start with the basic expression for a KL divergence for a probability space X, such that $\eta \epsilon X$;

$$D[q||p] = \sum_{\eta} q(\eta) log \frac{q(\eta)}{p(\eta)} \tag{1}$$

Simply put the KL divergence measures the difference between two probability distributions q and p. The (variational) free energy can be written in various ways as a mixture of a KL divergence and an expected energy.

Consider a partition of some system into states (η) that are external to an artefact, particle or person and particular states (π) that constitute the particle or person in question. These states can be further divided into internal states (μ) and blanket states (b) that intervene between the internal and external states.[1] With this partition in mind, we can now express the free energy as the expected surprisal or self-information of blanket states—given external states or their latent causes—plus a KL divergence between posterior beliefs and the (Bayesian) beliefs parameterised by internal states. Equivalently, we can express the free energy as the expected log likelihood of blanket states (e.g., sensory inputs) plus the KL divergence between Bayesian beliefs and prior beliefs about external states. In the free energy principle, these Bayesian beliefs are parameterised by internal states. This is denoted by the subscript in q_{μ}, in the following expression for free energy:

$$\begin{aligned} F(q,b) &= E_q[-\log p(b|\eta)] + D[q_{\mu}(\eta)|p(\eta)] > -\log p(b) \\ &= \text{inaccuracy} + \text{complexity} > \text{surprisal} \end{aligned} \tag{2}$$

The first two terms on the right-hand side of Eq. 2 correspond to inaccuracy and complexity respectively. In short, free energy provides an upper bound on self-information or surprisal, which can be read as scoring Bayesian beliefs in terms of their ability to explain external impressions on blanket states—e.g., sensory inputs or scientific measurements of some particle, agent or observer—as simply as possible. The inequality above means that minimising free energy minimises self-information and can be read as minimising complexity under accuracy constraints. The term 'complexity' scores the difference between Bayesian beliefs about latent or hidden causes (η) before and after is received some sensory evidence (b). The main insight of this paper is that Eq. 2 also provides a measure of realism. Surprisal is the upper limit on the subtotal of the internal inaccuracy and complexity of the self-organising particle. By respecting this upper limit, the organism makes itself a model of its environment.[2] In other words, it provides a reliable representation of its environment. Assuming that realism is mainly (if not totally) about having faithful enough

[1] In the context of the FEP, a Markov blanket is used to establish a statistical boundary that conditionally separates a system (such as, but not exclusively, a biological entity) from its external environment. It specifies which variables within the system are conditionally independent of the variables outside the blanket. Thus, blanket states refer to the internal states of a system that lie within its Markov blanket, and Markov blankets are used to specify the pertinent information required for making predictions or inferences about the system's internal states.

[2] One way of thinking about this is to note that self information is the negative of log evidence. Therefore, minimising self information or surprisal is the same as maximising the evidence for a probabilistic model of external states that is parameterised by internal states. This view is sometimes referred to as self evidencing.

representations of the environment, the present project offers to operationalise the reliability of representations (and thereby the notion of realism) in terms of finding the balance between inaccuracy and complexity (or the balance between accuracy and generality, on which more will be said later in this paper).

FEP, its general and neat formal articulation, and its capacity for representing biological facts have been at the centre of interesting philosophical debates (Kirchhoff et al., 2022; van Es & Hipolito, 2020). This paper aims to go beyond the extant discussions to explore the capacity of FEP for underpinning a new operational measure for realism. The notion of realism that we use here is derived from a Cognitive version of Structural Realism (CSR) [see Jones, (2019) for a critical review].

Scientific realist theories of various stripes aim to ground the empirical success of theories in their truth-likeness. Structural Realism (SR) is a version of realism that accounts for the cumulative growth of theories at the level of structure or form rather than content (Worrall, 1989). When elaborated metaphysically, CSR holds that structures are fundamental and individuals are derivative (French, 2014; Ladyman & Ross, 2007). Orthodox versions of SR, such as Steven French's (French, 2006; French & Ladyman, 1999) aim to model scientific representations via quasi-set theory and model theory (with partial structures), but according to CSR, this approach to representations is too abstract to contribute to reinforcing the realist core of SR (Beni, 2019a, 2019c chapter 3). According to this critique, to establish its claim to realism, SR must reinforce its account of structural representations with an intelligible account of how scientists (as actual biotic self-organising systems) can represent the structure of the world to themselves (both at the level of individuals and scientific communities) (Beni, 2018, 2021c). And CSR relies on the theoretical framework of FEP to construct its account of scientific representation. According to CSR, the information processing of the cognitive systems under the FEP underwrites the formation and verification of scientific theories.

> Informational structures of the scientific theories, which could be regimented by the unified entropy-based informational framework, latch onto reality on the basis of the predictive coding capacity of the brain. Thus we can account for the connection between the unified entropy-based informational framework (which regiments the informational structure of theories) and the world on the basis of the brain's capacity for decreasing the discrepancy between its models and reality (Beni, 2018, pp. 640–641).

Before going further, it should be noted that recent articulations of CSR do not assume that scientific inferences are exclusively grounded in the neurocomputational mechanisms of individual brains. Responding to Jones's (2019) worry about the negligence of social aspects of scientific practice, Beni (2021c) has argued that the cognitive structures that are at issue in CSR can and indeed do include patterns of distributed collective scientific knowledge. These patterns are distributed in scientific groups or even whole scientific communities. Drawing on preceding work of Giere (2002), Hutchins (1995), Kirsh (2010), and Nersessian (2003), the recent articulation of CSR indicates that cognitive embodied structures that are at issue in CSR could be construed along the lines of moderate versions of embodiment and enactivism. At some level (perhaps a basic one), scientific representations are

still taken to be formed in the brains, but brains are embodied and situated within ecological and social contexts. This provides a purchase for developing a distributed account of scientific representations in terms of patterns of human–human and human-artefact dynamics, which means the inferences that are at issue in scientific practice are not directly and exclusively supported by any individual scientist's brain. In short, CSR recognises that science is a socio-cultural practice. Scientists are organised into groups, and they use artefacts such as computers and external representations. The collective cognitive activity of individual scientists and its extension into external representations such as computational tools, laboratories, etc. are explicated by CSR as forms of adaptive complementary social behaviour under the rubric of FEP. This insight draws on Giere's account of the interrelatedness of the cognitive and social aspects of scientific activity, between which no sharp divide can be stipulated (see Beni, 2021c, p. 78). In this picture, scientific knowledge emerges from collective cognitive activity, which is affected by the social organisation and culturally evolved tools, as well as the cognitive abilities of individuals.

CSR's account of collective scientific knowledge (that is distributed into patterns of human–human and human-artefact dynamical interaction) is still supplemented by FEP. According to CSR, collective scientific practice, when taking place successfully, leads to the minimisation of the collective information entropy of the system as a whole. There is no water-tight proof to demonstrate that populations as large as laboratories or scientific groups are *deliberately* and *consciously* participating in minimising their collective entropy. In fact, even simpler accounts of adaptive complementary social behaviour under FEP require further experimental support. But at least the theoretical foundations of FEP can be trusted with the job of accounting for collective cognition in terms of minimising free energy by using shared generative models in different communities—from simple communities such as a population of neurons and a couple of birds that try to perform a duet based on the same sonic template to more sophisticated communities, such as scientific ones (Friston & Frith, 2015a, 2015b; Kirchhoff, 2018; Ramstead et al., 2019). I shall put my point in context by drawing on a working example that has been originally discussed by Chandrasekharan and Nersessian (2015).

Chandrasekharan and Nersessian (2015) argue that scientific cognition is distributed in the web of implicit sensorimotor processes of diverse contributors. To instantiate their view, they show how distributed information processing provides solutions to some concrete problems. For example, the problems of how to develop new drugs, build RNA folds, or explain how retinal cells detect motion can be handled by crowdsourcing them between multitudes of players in games such as *Foldit, EteRNA*, *Eyewire*. In this context, Chandrasekharan and Nersessian construe scientific cognition in terms of distributed processes; scientific representation is explicated in terms of building up external representations. I adapt their insight and below will put it in the context of a bioengineering lab.

Let us assume that A is a theoretical modeller and B is an experimentalist in the context of a bioengineering lab. Their collective aim is to construct external representations and fit them into internal ones. As Chandrasekharan and Nersessian remark, the task will be accomplished collectively, via cognitive processes that are distributed in a system that comprises the modeller, the experimenter,

and different artefacts—such as diagrams, graphs, papers, databases, search engines—that contribute to the accomplishment of the task. The collaboration is fruitful when A's and B's respective active inference schemes are coupled together and A and B collaborate on the basis of a shared generative model. I shall elaborate immediately. At the earliest stage of collaboration, A and B want to start to understand one another and get a grasp of the goals of the research. They want to figure out how the tasks are distributed and what are the available representational tools. In the beginning, there will be a rather high amount of prediction errors included in A's and B's respective models of what the other thinks. This means that, as Chandrasekharan and Nersessian's study indicates, at first, the collaboration between A and B is strained. This is because A's and B's respective models of each other are erroneous in the beginning. More precisely, as Chandrasekharan and Nersessian (2015, p. 1754) argue, at first the participants "had different representations of the mechanism, different levels of control, different goals/objectives, and little understanding of the nature of these differences". The divergence between A's and B's respective representations of mechanisms and levels of control could become too large to allow for efficient collaboration so much so that eventually the divergence may prevent them from achieving the goals of the research. For example, it might become the case that the divergence between perspectives is so big that B, who is the experimentalist, does not take the theoretical predictions of A, who is the modeller, seriously to try to test them carefully enough (Chandrasekharan & Nersessian, 2015, p. 1749). Moreover, even if A and B want to try to predict one another's intentions and actions without a shared basis, the problem of regress may raise its head. This is because A must be able to predict B, who aims to predict A, who aims to predict B, and so forth. The question is how A and B may succeed in constraining the discrepancy between their individualistic models and get a handle on intersubjective facts about the status of one another as well as goals, objectives and the involved mechanisms of their common research. It is obvious that for the research to succeed, A and B must begin to collaborate efficiently when their active inference schemes are coupled together. This means that they begin to subscribe to the same narrative so as to coordinate their respective active inference schemes. To enhance the efficiency of their collaboration, A and B must be able to minimise the discrepancy between their respective perspectives, say, when B the experimentalist begins to invest in the representations constructed by A the modeller. Thus, the minimisation of discrepancy happens when the collaboration starts to take off.

The example that has been used in the previous paragraph is inspired by Chandrasekharan and Nersessian's account. However, despite its merits, their account does not elaborate on the underpinning mechanisms of the formation of trust and collaboration. So, the main question is, what are the cognitive mechanisms of minimisation of uncertainty and the evolution of trust? The minimisation of the discrepancy between the respective perspectives can be explained best based on generalised synchrony under FEP. Under this model, the outcome of the prediction error

minimisation of one system could be predicted from the perspective of the other system based on a shared narrative or a shared generative mode (Beni, 2021c; Friston & Frith, 2015a). A and B would not be able to overcome their uncertainty about one another's goals, intentions, and the division of labour without getting entrapped into an infinite regress unless they can rely on a shared generative model that sets the stage for coordinating the division of responsibilities for fulfilling a given task under a social hierarchy superimposed on the structure of the distribution of the task.

The main point of this paper is to find a measure for realism in the context of CSR. CSR—which also accommodates an account of the dynamics of human–human and human–artefact integration in the context of scientific practice under the rubric of FEP—aspires to be a version of realism. Thus far, reasons that have been marshalled in favour of the realist core of CSR have been conjured from evolutionary biology (which fits the naturalist tendency of CSR). The present study aims to go further to find a measure for the notion of realism without presupposing the possibility of direct access to the causal structure of the world. The attempt at setting the measure for realism in terms of finding the balance between generality and accuracy under FEP is quite compatible with the distributed-collective take on scientific knowledge. This is because FEP underlies a viable neurocomputational account of social cognition. FEP theorists argue that mechanisms of minimising uncertainty and active inference underwrite computational models of social perception, social learning, social signalling and generally social inferences (Molapour et al., 2021). Minimisation of uncertainty in the context of social cognition, too, consists in putting surprisal as the upper bound on the subtotal of inaccuracy and complexity. This is explicable in terms of striking the balance between the accuracy and generality of models. If so, there will be formal consistency between using FEP to define a measure for scientific realism on the one hand and a social conception of scientific practice on the other, given that computational models of social cognition, too, can be specified in the formalism of minimising free energy. The main point is that from both an individualistic and collective perspective on science, there is no model-independent access to the unobservable parts of the world to confirm the veracity of representations (also see Beni, 2019a, chapters 3 and 4). Bearing that point in mind, we will endeavour to find a measure for realism from within the framework of CSR (in terms of the balance between accuracy and generality.

To recap, CSR gives up on some of the less modest claims of scientific realism—the conception of mind-independent reality is replaced with a perspectival, cognitive conception. However, this scientifically informed and modest version of realism would appeal to a radical naturalist. In the past few years, CSR was expanded to account for social aspects of scientific practice by relating FEP's view on alignment to dynamic systems theory (Beni, 2021b, 2021c) and explicate a naturalised account of scientific methodology and scientific inference that draws on FEP to address the issues of scientific model making and theory choice (Beni & Pietarinen, 2021; Pietarinen & Beni, 2021). What CSR still lacks is a clear sense of realism. In this paper,

we show how CSR can develop a perspective on scientific realism based on its naturalist account of selecting models.

3 Realism and Accuracy of Models

Levins's (1966) account of the strategy of model-based science in biology indicates that the complex process of scientific model-making is not simply a matter of setting up mathematical models that can provide "a faithful, one-to-one reflection of this complexity" (Levins, 1966, p. 422). The central insight of Levins's paper is that generality, realism, and precision about the goals of understanding, predicting and controlling the world cannot be simultaneously maximised (Levins, 1966, p. 423). Absent maximising all three factors simultaneously, Levins (1966, p. 423) conceives of three approaches for dealing with generality, realism and precision.

I. The modeller sacrifices generality for realism and precision. The focus would be on the short-term behaviour of the organism in particular situations. This paper is not concerned with (I) and does not engage in an in-depth discussion of it.
II. The modeller sacrifices realism to generality and precision, more or less in the same way that in physics, physicists construct models of perfect gases or frictionless planes.
III. The modeller sacrifices precision to generality and realism, caring for the long-run qualitative rather than quantitative results.

In his discussion of these three approaches, Levins identifies realism with accuracy alone and does not assume that generality and precision can contribute to realism as well (Beni, 2022). This point becomes prominent in Levins's negative evaluation of the second strategy, where (1966, p. 422) he gives examples of the kinetic theory of gases and frictionless planes and suggests that the involved models (such as general equations in the kinetic theory) are unrealistic in the sense that they lack accuracy.

According to Levins, the second strategy, namely (II), leads to the minimisation of realism in the sense that involved models do not provide accurate enough representations of the causal structure. He remarks that this lack of accuracy—which is a result of the generalisation of models—is prevalent in models of physics, which abstract away from full details and idealise them—hence the reference to the kinetic theory or the Galilean account of motion on inclined planes. Whether or not the distortion that is caused by idealisation undermines the realist understanding of theories of physics is an important question (Cartwright, 1983; Frigg, 2010; Suárez, 1999). However, we do not need to engage in that fundamental discussion to develop

our account here. For, the assumption that we make is quite minimal. We assume that, from the fact that scientific practice relies on idealisation, approximation, application of ceteris paribus laws, and so on, *it does not follow* that the class of models that instantiate theories fail to provide precise/faithful (enough) representation of their target systems (McMullin, 1985).[3] In the same vein, we assume that adding *unnecessary* details to models is always unwelcome in both special sciences and physics (Craver & Kaplan, 2018). We do not think these assumptions are demanding. It is some kind of truism to say that more details are not always better, and adding details marked as unnecessary will not be helpful. The more important question, that needs to be articulated and addressed with more care is how to determine what details are necessary and what details are not (Beni, 2022). This is the question that we will discuss in the remainder of this paper. We will argue that CSR can draw on FEP once more to provide a realist measure of how much detail should be retained without undermining the representational capacity of theories.

I shall clarify my point with an example before going further. When reproving the use of generic equations in his section on strategy, Levins remarks on the models of the Volterra predator–prey equation. This equation glosses over physiological details as well as the effect of a species' population density on its rate of increase. However the realist tendency of the Volterra equation can be defended as well on the same logic that McMullin used to deal with the implied idealisation of the kinetic theory (see the previous paragraph). Despite glossing over some physiological details, the class of models under the Volterra equations[4] still manages to achieve empirical adequacy. The main point here is that although the Volterra equation representations are not fully detailed and complete, there is no reason to think that they do not provide truthful representations of the general relationship between prey and predators. We do not of course assume that the equations need to refer to the mind-independent structure of reality. The heart of the enquiry is elsewhere.

It is true that Volterra's model misses some details, for example, about the physiological constitution of prey and predators. Nevertheless, the model comes with non-negligible explanatory and predictive power. We do not submit that the explanatory power of the model is grounded in the mind-independent causal structure of the world. Our question is about how much accuracy is enough for bolstering realism

[3] As Levins remarks, the kinetic theory of gases does not specify the internal structure of molecules that are taken as constituents of the gases. So, the models of kinetic theory idealise some concrete facts about the intrinsic nature or structure of the molecules. But not all facts about the internal structure of the molecules contribute to increasing the explanatory or predictive power of the theory. Maximising accuracy about all the details (or having complete models) undermines the representational power of the theory, and as such it could hardly maximise realism.

[4] Assuming that V is the size of the prey population and P denotes the size of predator population, r is the rate of the growth of the prey population and m is the death rate of predators, Volterra's model of fluctuation of predator–prey dynamics (Volterra, 1926) holds that:

$$\frac{dV}{dt} = rV - (aV)P$$

$$\frac{dP}{dt} = b(aV)P - mP$$

about models. In the present case, more details about the physiological constitution of the prey and predator as well as the properties of the environment—whether it contains, say, saltwater or brackish—would increase the accuracy of a given application of Volterra's model, but these details do not necessarily contribute to providing a more realistic explanation of why cessation of fishing led to the decrease in fish population, in defiance of the earlier expectations (Beni, 2022). In fact, as Weisberg remarks, it is not always possible (or even desirable) to have complete models of target systems in their full complexity, especially when one strives to model complex systems (Weisberg, 2007, p. 626).

In short, we deny that realism entails the maximal accuracy of models with respect to their specified applications. There is no doubt that realism about models mandates concern for some degree of accuracy of models. Our question is about how to find the right balance between the accuracy and generality of models that could be the loci of realist commitments.

4 Free Energy Principle as the Measure for Realism

Thus far in this paper, I have argued that realism should not be specified in terms of the accuracy of models alone. Instead, the paper proposes to reconceptualise realism in terms of the balance between the generality and accuracy of models. From a scientific realist perspective, models should be accurate enough to impart detailed enough information about the properties of the target systems, but at the same time, they must be generic enough to represent only properties that are relevant to specific applications (and possibly, explain how these properties are connected with the physical foundations of the world). In other words, models that are over-parameterised can provide a very accurate fit and match their target system very well. Realism is about finding the right balance between accuracy and generality. This provides a nice operational notion of realism in the context of CSR.

Being a version of structural realism, CSR seeks to ground the veracity of scientific representation in the (individual and collective) ability of scientists (as biotic self-organising systems) to represent the causal structure of the world to themselves, and thereby it links the veracity of scientific representations with the fact that in order to maximise the chance of their survival, biotic self-organising systems need to fasten their cognitive grip on a non-negligible portion of the causal structure of the real world (Beni, 2019a, chapter 6). In this paper thus far realism has been operationalized in terms of striking the balance between generality and accuracy. I finish this paper by briefly spelling out this measure in terms of FEP. The same free energy minimising mechanisms that natural self-organising systems apply to stay in non-equilibrium steady states enable the notion of realism as the balance between accuracy and generality. This move will be in harmony with CSR's radically naturalist enterprise to ground scientific representation in the dynamical interplay between the organism and its (social and biological) environment under FEP (more on this later).

FEP itself seems to owe part of its success to the fact that the right combination of generality and accuracy is incorporated into its formulation. Relying on generic

equations that model the minimisation of free energy, FEP draws unificatory links between various fields in psychology, life science, computational neuroscience, and even social cognition (Friston, 2010; Hesp et al., 2019; Vasil et al., 2020). At the same time, by invoking detailed enough causal models of the neurophysiological mechanisms of cognition (Büchel & Friston, 1997; Friston et al., 2003) it provides rather accurate accounts of embodied mechanisms of cognition and life at the level of neuronal nested populations (Kirchhoff & Kiverstein, 2019; Pezzulo et al.,, 2017). So, FEP itself seems to be articulated by finding the right balance between generality and accuracy. FEP not only owes its success at least partly to striking the right balance between generality and accuracy, it is also an enabler of the harmony between accuracy and generality of scientific models, where this harmony is the main constituent of scientific realism.

When introducing FEP in Section 2 of this paper, I remarked that KL-divergence could provide a measure of relative entropy as a centrepiece of FEP. For the sake of simplicity, I draw on a simpler version of KL-divergence (Eq. 1), which is also used in Mann et al.'s (2021) articulation of FEP, to show how FEP can optimise the balance between accuracy and generality of models. I shall flesh out my proposal immediately.

Let us express complexity in terms of prior and posterior beliefs about some latent causes or hypothesis $D[q_{\mu}(\eta)|p(\eta)]$, as in Eq. 3. We assume that mechanisms of formation of *scientific* hypotheses and testing them are grounded in mechanisms of minimisation of free energy under FEP [*a la* (Beni, 2019a chapters 6 and 7; Beni & Pietarinen, 2021; Pietarinen & Beni, 2021)]. To press our point, instead of construing 3 as a formal basis for developing an account of cognition per se, we interpret Eq. 3 to represent the mechanisms of constructing and testing scientific hypotheses, which, according to CSR, are grounded in neurophysiological mechanisms of minimising prediction error under FEP. The equation, when interpreted in terms of generative models that underwrite cognition per se (e.g., in the case of Eq. 1) (originally) articulates the relation between prior predictions embedded in the generative models and posteriors about hidden states, but we expand its interpretation by submitting that $p(\eta)$ represents the *scientific* model's prior hypotheses about the features of the territory, and $q(\eta)$ denotes posterior beliefs about the features of the scientific target system (which is supposed to be represented by the class of models). Parts of the target systems are represented by scientific hypotheses. When seen in this light, FEP underwrites a likely story of how natural phenomena are represented by scientific models that can be updated against the evidence. Under this interpretation, w denotes a specific configuration of worldly states, and let $p(\eta)$ denote the prior representational states of a given scientific hypothesis about external state of affairs η and $q(\eta)$ denote the posterior beliefs about the same states after imputing the model to the target system and updating the hypothesis.

The interpretation of Eq. 3 in the previous paragraph just submits a technical articulation of the main insights behind CSR, as previously presented in (Beni, 2018, 2019a). Now we go further to unpack the implications of this proposal for Levins's notion of generality, and indeed for the definition of scientific realism in general. Our main insight here is that if the models are completely accurate, then the divergence between the two sets of distributions p and q will, in general, be high–in

order to fit the data at hand. This means that the model closely fits the target system completely. The downside of this is that models would overfit the target system in a way that makes them fail to serve in the explanation or prediction of anything aside from the specific configuration of η. Complete accuracy is not a desirable feature of *scientific* models because it makes them a bit too local to serve their scientific representation purpose.

In order to support realism about scientific models, we not only must avoid maximal accuracy but also over-generalisation. This is because models that dismiss all relevant details fail to accomplish their representational task as well. This is tantamount to saying that models that we cannot use to form any accurate predictions about the configuration of any worldly state would fail to serve their scientific purpose. In that circumstance, any occurrence would become surprising for the agent, because the agent's hypotheses or models would become too general to make any precise predictions about what would happen next. So, let η represent a specific configuration of worldly state and b denote a piece of evidence that the agent has for supporting the hypothesis about η. In this circumstance, $p(b, \eta)$ represents a probabilistic generative model about external or worldly states (η) and the empirical evidence for that hypothesis (b). In this setting, the accuracy $E_q[\log p(b|\eta)]$ is given by Eq. 3, where, $p(b|\eta)$ represents the likelihood of the evidence given the hypothesis about the worldly state. The equation indicates how surprised we would be in case of divergence between our prior beliefs about evidence being the case given the truth of the hypothesis about the state of the world on the one hand and q (η) or our posterior belief after imputing the model to the target system on the other. In case the model is too general, the divergence between $q(\eta)$ and $p(b)$ would exceed expectations. That is to say beliefs about the world would be so general that they could not be used to predict any specific distribution of external states of with any accuracy and thus any future states would be surprising to the agent that uses the over-generalised predictive models. Unless the divergence between $q(\eta)$ and $p(\eta)$ could be constrained, the model would fail to serve its explanatory and predictive goals. Under FEP, the evolutionary cost of the failure is that models that maximise surprise would minimise survival. In the context of scientific practice, the cost of using over-generalised models is the falsification of theories.

Thus far in this section, we constructed the operational measures of the accuracy and generality of models. Now we construct a measure for evincing the relationship between accuracy and generality. (the solution is inspired by Mann et al., (2021) who use this (or a very similar) approach to find the balance between over-fitting and the failure to explain);

$$M(q, b) = E_q[\log p(b|\,\eta)] - D[q_\mu(\eta)\,|\,p(\eta)] \;<\; \log p(b) = \text{accuracy} - \text{complexity} \tag{3}$$

M is the measure for realism, saying that a class of models whose application conforms to general guidelines of realism must avoid both extremes of being too

accurate and too general. This measure just is the negative variational free energy in Eq. 2.[5] The first term represents the accuracy, whereas the second term represents the complexity penalty for too much generality. By constraining both accuracy and generality, the FEP provides a measure of realism.

Admittedly, it is unsurprising that Eq. 3 can apply to the scientific context. When stating their formulation of FEP, authors sometimes refer to the scientific context metaphorically (Hohwy, 2013; Mann et al., 2021). What is interesting in this context though is to go beyond mere metaphors and admit that scientific practice (both at the individual and collective levels) is based on the neuro-computational mechanisms of the minimisation of free energy, and to admit that the general realist perspective could be grounded on the cognitive mechanisms of model-selection. CSR aspired to develop such a cognitive perspective of science by drawing on FEP. Using FEP as a measure of realism—in terms of the harmony between accuracy and generality—amends CSR's lack of a clear statement on realism.

5 Concluding Remarks

The paper aimed to develop a cognitive-operational measure of scientific realism. We started by revisiting Richard Levins's identification of realism with the accuracy of scientific models in life sciences and argued that realism cannot be specified in terms of accuracy alone but as the harmony between accuracy and generality. To put flesh on this skeletal definition of realism (as the balance between accuracy and generality) we fell back on a cognitive version of structural realism that grounded its account of scientific representation in the minimisation of free energy under the free energy principle (FEP). Crucially, the FEP can be used to articulate an operational measure of finding the right balance between generality and accuracy. This proposal is in line with cognitive structural realism (CSR), which submits that scientific practice is a finessed form of the capacity of self-organising systems to minimise the discrepancy between their internal models and the causal structure of reality. At the same time, the enterprise of this paper provided CSR with a congenial operational definition of realism. In short, FEP has been used to account for various theories about self-organising systems and their sentient behaviour; CSR projects this account to the domain of scientific collective knowledge. The upshot is that FEP does not need to be used to argue for or against scientific realism but, rather, can be regarded as itself furnishing a theory of scientific realism.

Acknowledgements I am grateful for the constructive comments provided by the anonymous referees of this journal. Additionally, I would like to express my gratitude to Karl Friston for reviewing the formalism of the paper and providing valuable revisions

[5] In machine learning, this is known as an evidence lower bound (ELBO) because the sign of the variational free energy has be reversed. In other words, to maximise realism, one can maximise the lower bound on log evidence.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK).

Declarations

Conflict of interest There are no financial or non-financial interests that are directly or indirectly related to this submission.

References

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology and Philosophy*. https://doi.org/10.1007/S10539-021-09807-0

Beni, M. D. (2016). Epistemic informational structural realism. *Minds and Machines, 26*(4), 323–339. https://doi.org/10.1007/s11023-016-9403-4

Beni, M. D. (2018). Syntactical informational structural realism. *Minds and Machines, 28*(4), 623–643. https://doi.org/10.1007/s11023-018-9463-8

Beni, M. D. (2019a). *Cognitive structural realism: A radical solution to the problem of scientific representation*. Springer.

Beni, M. D. (2019b). Conjuring cognitive structures: Towards a unified model of cognition. In A. Nepomuceno-Fernández, L. Magnani, F. Salguero-Lamillar, C. Barés-Gómez, & M. Fontaine (Eds.), *Model-based reasoning in science and technology MBR 2018* (pp. 153–172). Springer.

Beni, M. D. (2019c). The benacerraf problem as a challenge for ontic structural realism†. *Philosophia Mathematica*. https://doi.org/10.1093/philmat/nkz022

Beni, M. D. (2021a). A critical analysis of Markovian monism. *Synthese, 199*(3), 6407–6427. https://doi.org/10.1007/S11229-021-03075-X

Beni, M. D. (2021b). Cognitive penetration and cognitive realism. *Episteme*. https://doi.org/10.1017/EPI.2021.39

Beni, M. D. (2021c). Inflating the social aspects of cognitive structural realism. *European Journal for Philosophy of Science, 11*(3), 1–18. https://doi.org/10.1007/S13194-021-00401-5

Beni, M. D. (2022). Dosis sola facit venenum: Reconceptualising biological realism. *Biology & Philosophy, 37*(6), 1–18. https://doi.org/10.1007/S10539-022-09884-9

Beni, M. D., & Pietarinen, A.-V. (2021). Aligning the free-energy principle with Peirce's logic of science and economy of research. *European Journal for Philosophy of Science, 11*(3), 1–21. https://doi.org/10.1007/S13194-021-00408-Y

Büchel, C., & Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex, 7*(8), 768–778. https://doi.org/10.1093/CERCOR/7.8.768

Cartwright, N. (1983). How the laws of physics lie. *Oxford University Press*. https://doi.org/10.1093/0198247044.001.0001

Chandrasekharan, S., & Nersessian, N. J. (2015). Building cognition: The construction of computational representations for scientific discovery. *Cognitive Science, 39*(8), 1727–1763. https://doi.org/10.1111/cogs.12203

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy, 36*(5), 1–26. https://doi.org/10.1007/S10539-021-09818-X

Craver, C. F., & Kaplan, D. M. (2018). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy015
Floridi, L. (2008). A defence of informational structural realism. *Synthese, 161*(2), 219–253. https://doi.org/10.1007/s11229-007-9163-z
Floridi, L. (2009). Against digital ontology. *Synthese, 168*(1), 151–178. https://doi.org/10.1007/s11229-008-9334-6
French, S. (2006). Structure as a weapon of the realist. *Proceedings of the Aristotelian Society (hardback), 106*(1), 170–187. https://doi.org/10.1111/j.1467-9264.2006.00143.x
French, S. (2014). *The structure of the world: Metaphysics and representation*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199684847.001.0001
French, S., & Ladyman, J. (1999). Reinflating the semantic approach. *International Studies in the Philosophy of Science, 13*(2), 103–121. https://doi.org/10.1080/02698599908573612
Frigg, R. (2010). Models and fiction. *Synthese, 172*(2), 251–268. https://doi.org/10.1007/s11229-009-9505-0
Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787
Friston, K. J., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics, 102*(3), 227–260. https://doi.org/10.1007/s00422-010-0364-z
Friston, K. J., & Frith, C. (2015a). A duet for one. *Consciousness and Cognition, 36*, 390–405. https://doi.org/10.1016/J.CONCOG.2014.12.003
Friston, K. J., & Frith, C. D. (2015b). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143. https://doi.org/10.1016/j.cortex.2015.03.025
Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage, 19*(4), 1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7
Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy, 22*(5), 516. https://doi.org/10.3390/E22050516
Giere, R. N. (2002). Scientific cognition as distributed cognition. In P. Carruthers, S. Stich, & M. Siegal (Eds.), *The cognitive basis of science* (pp. 285–299). Cambridge University Press.
Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M. & Friston, K. (2019). *A multi-scale view of the emergent complexity of life*: *A free-energy proposal*. *Springer Proceedings in Complexity* (pp.195–227). https://doi.org/10.1007/978-3-030-00075-2_7
Hohwy, J. (2013). *The predictive mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199682737.001.0001
Hutchins, E. (1995). Cognition in the wild. *The MIT Press*. Retrieved from https://mitpress.mit.edu/books/cognition-wild
Jones, M. (2019). Cognitive structural realism: A radical solution to the problem of scientific representation. *Philosophical Psychology, 33*(5), 772–775. https://doi.org/10.1080/09515089.2020.1765327
Kirchhoff, M. (2018). Hierarchical Markov blankets and adaptive active inference: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews*. https://doi.org/10.1016/J.PLREV.2017.12.009
Kirchhoff, M. D., & Kiverstein, J. (2019). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*. https://doi.org/10.1007/s11229-019-02370-y
Kirchhoff, M. D., Kiverstein, J., & Robertson, I. (2022). The literalist & fallacy the free energy principle: Model-building scientific realism and instrumentalism. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/720861
Kirsh, D. (2010). Thinking with external representations. *AI and Society, 25*(4), 441–454. https://doi.org/10.1007/s00146-010-0272-8
Ladyman, J., & Ross, D. (2007). *Every thing must go*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199276196.001.0001
Levins, R. (1966). The strategy of model building in population biology. *American Scientist, 54*(4), 421–431.
Mann, S. F., Pain, R. A., & Kirchhoff, M. (2021). Free energy: A user's guide. *Biology & Philosophy, 37*(4), 33.
McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A, 16*(3), 247–273. https://doi.org/10.1016/0039-3681(85)90003-2
Molapour, T., Hagan, C. C., Silston, B., Wu, H., Ramstead, M., Friston, K., & Mobbs, D. (2021). Seven computations of the social brain. *Social Cognitive and Affective Neuroscience, 16*(8), 745–760. https://doi.org/10.1093/SCAN/NSAB024

Nersessian, N. J. (2003). Interpreting scientific and engineering practices: Integrating the cognitive, social, and cultural dimensions. In M. E. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.), *Scientific and technological thinking* (pp. 17–56). Lawrence Erlbaum Associates Publishers.

Pezzulo, G., & Levin, M. (2017). Embodying Markov blankets: Comment on "Answering Schrödinger's question: A free-energy formulation" by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews*. https://doi.org/10.1016/J.PLREV.2017.11.020

Pietarinen, A.-V., & Beni, M. D. (2021). Active inference and abduction. *Biosemiotics*. https://doi.org/10.1007/s12304-021-09432-0

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*. https://doi.org/10.1016/J.PLREV.2017.09.001

Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy, 22*(8), 889. https://doi.org/10.3390/e22080889

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*. https://doi.org/10.1177/1059712319862774

Suárez, M. (1999). Theories, models, and representations. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 75–83). Springer.

van Es T, Hipolito I (2020). Free-energy principle, computationalism and realism: A tragedy. http://philsci-archive.pitt.edu/18497/

Vasil, J., Badcock, P. B., Constant, A., Friston, K., & Ramstead, M. J. D. (2020). A world unto itself: Human communication as active inference. *Frontiers in Psychology, 11*, 417. https://doi.org/10.3389/fpsyg.2020.00417

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature, 118*(2972), 558–560. https://doi.org/10.1038/118558a0

Weisberg, M. (2007). Forty years of 'The Strategy': Levins on model building and idealization. *Biology & Philosophy, 21*(5), 623–645. https://doi.org/10.1007/s10539-006-9051-9

Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica, 43*(1–2), 99–124. https://doi.org/10.1111/j.1746-8361.1989.tb00933.x