Biology & Philosophy (2022) 37: 34
https://doi.org/10.1007/s10539-022-09868-9

Teleosemantics and the free energy principle

Stephen Francis Mann[1,2] · **Ross Pain**[2]

Received: 15 September 2021 / Accepted: 12 July 2022 / Published online: 27 July 2022

Abstract
The free energy principle is notoriously difficult to understand. In this paper, we relate the principle to a framework that philosophers of biology are familiar with: Ruth Millikan's teleosemantics. We argue that: (i) systems that minimise free energy are systems with a proper function; and (ii) Karl Friston's notion of implicit modelling can be understood in terms of Millikan's notion of mapping relations. Our analysis reveals some surprising formal similarities between the two frameworks, and suggests interesting lines of future research. We hope this will aid further philosophical evaluation of the free energy principle.

Keywords Teleosemantics · The free energy principle · Active inference · Proper functions · Markov blankets

Introduction

Proponents of the free energy principle have ambitious goals. In its strongest formulations, the imperative to minimize free energy is claimed to provide a unified framework for understanding processes of life, mind, and even culture (Friston 2009, 2010, 2013; Kirchhoff 2018; Kirchhoff et al. 2018; Veissière et al. 2020; Rubin et al. 2020).

S. F. Mann was supported in part by an ANU University Research Scholarship, and the Australian Research Council under Laureate Fellowship Grant FL130100141. Ross Pain was supported by an ANU University Research Scholarship, the Australian Research Council under Laureate Fellowship Grant FL130100141, and the ANU Futures Scheme.

✉ Stephen Francis Mann
stephen_mann@eva.mpg.de; stephenfmann@gmail.com

Ross Pain
ross.pain@anu.edu.au

[1] Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

[2] Present Address: School of Philosophy, Australian National University, Level 6, 146 Ellery Crescent, Canberra, ACT 0200, Australia

These proponents argue that living systems are able to maintain stable boundaries by minimizing the unexpectedness of their sensory states. This is achieved by using active states to influence sensory states via manipulation of the world and of the organism's place in it. According to the view, it follows that organisms will appear to 'model' circumstances beyond their boundaries (Friston 2013). This allows a living system to remain within a restricted range of 'attracting states'; states which are beneficial to its ongoing success and which collectively make up its phenotype. In rough terms, this is the free energy principle.

So a central commitment of the free energy principle is the view that life, at multiple scales of organization, can be usefully interpreted as *implicitly modelling its environment*. Explaining what it means to implicitly model an environment, and how living systems might be interpreted as doing so, will be one of the core tasks of this article.

The free energy principle remains on the fringes of mainstream philosophy of biology.[1] This probably reflects long-standing concerns regarding the application of information theory, entropy and thermodynamics to biology, and, more broadly, scepticism with respect to the explanatory purchase of grand unified theories (Levins 1966; Morowitz 1986; Weisberg 2006). We take these concerns to be a serious challenge–perhaps *the* most serious challenge—facing proponents of the free energy principle; but for the purposes of this article, we sideline them. Our aim is to investigate the principle at a simpler level of analysis.

There is another reason for the free energy principle's fringe status. It is expressed using complicated mathematics and often obscure terminology. For instance, the 'energy' referenced by the free energy principle is not energy in the standard sense. Rather, it is an information-theoretic term. Calling the free energy principle FEP, Colombo and Wright (2018) outline the situation as follows:

> FEP's epistemic status remains opaque, along with its exact role in biological and neuroscientific theorizing. Conspiring against its accessibility are the varying formalisms and formulations of FEP, the changing scope of application, reliance on undefined terms and stipulative definitions, and the lack of clarity in the logical structure of the reasoning leading to FEP.
> Colombo and Wright (2018, p. 2)

Our goal in this article is to make some progress on this issue. We do so by combining the free energy principle with a framework that philosophers of biology are familiar with: teleosemantics. This compare and contrast exercise produces some interesting results. In particular, we suggest a correspondence between Karl Friston's conditions under which a system will minimize free energy, and Ruth Millikan's conditions under which a system will possess a direct proper function. Any system that ends up persisting in a non-equilibrium steady state via active inference is thus a system that has a proper function. Put more colloquially, minimizing free

[1] At the time of writing, the only articles mentioning the free energy principle in *Biology & Philosophy* are associated with this Topical Collection. It has received more attention from philosophers of cognitive science (Sprevak 2020; Williams 2021; Hohwy 2020; Bruineberg et al. 2021).

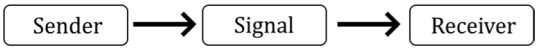

Fig. 1 The causal chain at the heart of the basic teleosemantic model. A *Sender* produces an intermediary, here labelled *Signal*, on which a *Receiver* conditions its behaviour. Sender and receiver must cooperate, which is analysed in terms of sharing a proper function

energy gets you proper functions. We can then begin to understand the claim that a system implicitly models its environment in terms of Millikan's notions of mapping relations between signals and the environmental circumstances they signify. Our goals are modest: the central motivation for integrating the two frameworks is to aid further philosophical evaluation of the free energy principle. But our analysis also points to important lines of future research.

We proceed as follows. "Teleosemantics and proper functions" section provides a brief primer on teleosemantics and the theory of proper functions. "Proper functions and free energy minimization" section outlines the structural similarities between the free energy principle and proper function. "Signals as internal models" section suggests a correspondence between semantic content in simple signals and the sense in which systems that minimize free energy can be said to harbour models of their environments. "Future research and closing remarks" section outlines some opportunities for further research.

Teleosemantics and proper functions

The literature on teleosemantics is extensive. Here our focus is Millikan's theory of sender–receiver teleosemantics, also known as biosemantics (Millikan 1989). Signals and proper functions play a central role in her account.

Signals

Teleosemantics defines a signal as an intermediary between a pair of cooperating devices: (1) a *sender*, which produces the intermediary either by performing a behaviour or emitting a physical item; and (2) a *receiver*, which conditions its own behaviour on the intermediary.[2] The sender-intermediary-receiver triad is a causal chain (Fig. 1). A key feature that differentiates teleosemantics from older causal accounts of mental content is that sender and receiver must have *proper functions*, and must be cooperating. A proper function is a causally downstream outcome that a device is designed to bring about. We will add more detail to this brief characterisation in

[2] Teleosemantics is usually described as a theory of representation; we use 'signal' to emphasize the liberality of the conditions by which the theory attributes representational status. Other versions of teleosemantics reject the requirement that representations have sender–receiver structure (Neander 2017; Shea 2018); it would presumably be misleading to substitute 'representation' for 'signal' in a discussion of those theories.

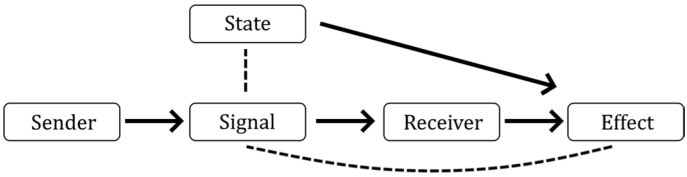

Fig. 2 The basic teleosemantic model. The *Receiver* has a proper function to bring about some *Effect* (in a causal model, this function would be specified as a requirement to set the effect variable to a certain value). However, an external *State* also has causal influence over the effect. The receiver *cannot* directly condition its behaviour on the value of this state. The *Sender*, which has as a proper function to help the receiver achieve its function, produces a *Signal* on which the receiver *can* condition its behaviour. Teleosemantics asserts that when the receiver conditions its behaviour on the signal and is more successful than it would have been otherwise, this increased success can only be fully explained by adverting to a relation between the signal and the state (upper dashed line). This relation is then the basic representational relation, or descriptive content. The signal also has directive content, interpreted as a command to bring about the required value of the effect variable (lower dashed line). Adapted from Millikan (2004, fig. 6.3, p. 78)

a moment. First we describe the kind of cooperation sender and receiver engage in, and why this merits treating the intermediary as a contentful signal.

The receiver has a function to perform, a downstream causal effect it is supposed to bring about. In the paradigm case, its success is dependent in part on an external circumstance which the receiver cannot directly observe. Conditioning its behaviour on the intermediary leads to greater success than acting unconditionally. The core commitment of teleosemantics is that explaining this improved success requires positing a relation between the intermediary and the external success-relevant circumstance. According to the theory, this relation is the basic form of semantic content. The intermediary is a signalling vehicle (or just 'signal', labelled Signal in the figures), and the external circumstance is its *descriptive content* (labelled State); see Fig. 2. Signals in simple models like this also have *directive content*, which is intuitively characterised as a command to bring about the required value of the effect variable (lower dashed line in Fig. 2).

The basic teleosemantic model depicted in Fig. 2 can be applied to phenomena throughout biology and cognitive science. Cells emit chemical messages that help coordinate and control joint behaviour. Social animals perform overt behaviours such as audible calls to assist conspecifics in finding food or avoiding predators. Nerve endings transmit electrical pulses via the central nervous system to the motor cortex, prompting a reflexive muscular response that protects the body from harm. These kinds of situations are very often described in terms of signalling, messaging, information or representation, and practitioners in those fields often draw on these concepts in giving explanations. One motivation behind teleosemantics is the promise of a general-purpose model that legitimises these explanatory practices.[3]

[3] There are several live issues in the teleosemantic literature that we ignore in this paper. Perhaps the most significant is how the basic model applies when sender and receiver do not cooperate, or do not cooperate perfectly. The cases discussed in "Proper functions and free energy minimization" and "Signals as internal models" sections include senders and receivers within single organisms, suggesting they will be fully cooperative. There are certainly times when components of organisms do not cooperate fully, but our assumption of full cooperation is an idealisation required to get the story off the ground.

Proper functions

There is a clear sense in which many biological devices have a function. They are adaptations, having selected effects that contribute to their proliferation. For example, the mammalian heart has a selected effect to pump oxygenated blood around the body. Hearts that achieve this effect contribute to survival and reproduction of the genes that produced them, thereby contributing to the production of more hearts in future. A common pattern of biological explanation follows the logic of selection, explaining a device's structure and behaviour by referring to the success of previous instances of the device. This is adaptationist, or more broadly selectionist, explanation.

The teleosemantic term for a selected effect is a *proper function*. Proper functions are not restricted to items produced by genes that proliferate due to natural selection acting on genetic lineages. Any device that owes its structure and/or dispositions to selection on the effects of certain 'ancestors' has a proper function. For example, a cognitive capacity in a trained rat to press a lever to retrieve food has lever-pressing as a proper function. A lever-pressing disposition has been reinforced (and alternative behaviours perhaps inhibited) by provision of food. The disposition 'proliferates' because previous manifestations of that disposition were followed by consumption of food. For a disposition to proliferate here means being more likely to occur than other possible dispositions. The period of selection—which in this case is a period of reinforcement—is confined to a single organism. It is nevertheless selection in the appropriate sense, because it is a process of differential retention of a certain disposition (lever-pressing) over others. In this case, the 'ancestors' of present lever-pressing behaviour are earlier instances of lever-pressing performed by the same rat.[4]

As a result, different kinds of selection process can give rise to the cooperative system depicted in Fig. 2. Even if a sender–receiver system is not a product of genetic selection, it may nonetheless have appropriate functions that recommend treating the intermediary as a contentful signal.

[4] The question of the relationship between learning processes and natural selection has been much discussed and is, we assume, not settled (Skinner 1981; Baigrie 1989; Catania 1999; Hull et al. 2001; Kingsbury 2008; Artiga 2010; Watson and Szathmáry 2016). All teleosemantics requires is that there is some explanatorily relevant similarity in the processes that give rise to functional behaviours. Millikan (1984, §§1–2) defends this claim extensively in giving the definition of proper function. We recapitulate key aspects of that definition in "Proper functions and free energy minimization" section.

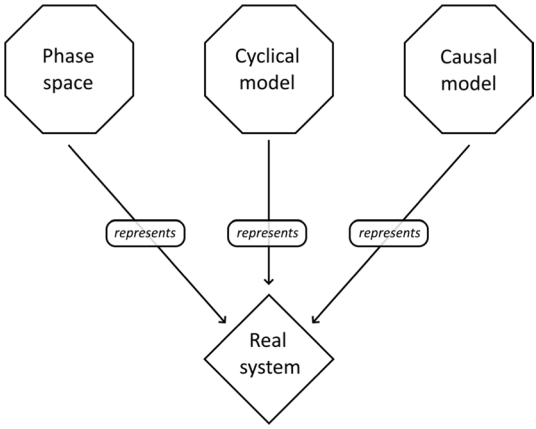

Fig. 3 Three ways to model a biological system. Causal models provide a definition of proper function. Phase spaces enable specification of the free energy principle. Cyclical models bridge the gap between them

Proper functions and free energy minimization

The argument

With the basic framework of teleosemantics on the table, we can consider its relationship to the free energy principle. We suggest a satisfaction relation between two sets of conditions:

> The conditions under which the free energy principle holds for a system satisfy (i.e. are a subset of)
> The conditions under which a system with a Markov blanket has a proper function (we will introduce the concept of a Markov blanket shortly).

By 'system' we are referring primarily to *models* of real systems. We formulate the definition of proper function within a *causal modelling* framework. The free energy principle is typically formulated within a dynamic modelling framework, depicted with *cyclical models* and *phase space* diagrams, all of which we will introduce shortly. Either kind of model can be used to represent a real system (Fig. 3). We aim to show that the two sets of conditions coincide *with respect to such models*. There will of course be further questions about which real systems (if any) those models faithfully depict. Proponents of active inference hope that their models capture observable features of real systems, and hence that the framework has empirical purchase. Our task here is not to evaluate these hopes directly, but to lay out some tools which might aid such an evaluation.

To show that the two sets of conditions are related in the proposed way, we must introduce: (i) Markov blankets; (ii) the conditions under which the free energy principle holds; and (iii) the conditions under which a system with a Markov blanket has a proper function. We first introduce Markov blankets and the free energy principle, before demonstrating how proper function relates to both.

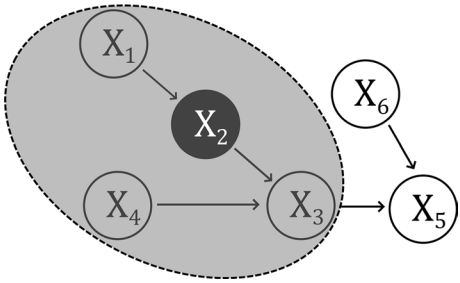

Fig. 4 A Markov blanket in a Bayesian network. The network represents the fact that a joint probability distribution $p(x_1, x_2, x_3, x_4, x_5, x_6)$ factorises in the following way: $p(x_1)p(x_4)p(x_6)p(x_2|x_1)p(x_3|x_2, x_4)p(x_5|x_3, x_6)$. The Markov blanket of a focal node such as X_2 consists of all nodes with respect to which X_2 is conditionally independent of every other node. With respect to X_1, X_3 and X_4, the focal node is conditionally independent of all other nodes (i.e. X_5 and X_6). Therefore, the Markov blanket of X_2 is the set $\{X_1, X_3, X_4\}$

Markov blankets

By the 1980s, statistical modelling techniques were being applied to increasingly complex phenomena. Models were apt to include many different variables, all of which could be related to each other in complicated probabilistic relationships. As a means of depicting these relationships succinctly, a kind of diagram called a *Bayesian network* was introduced. Bayesian networks depict variables as nodes, and probabilistic relationships between variables as lines drawn between nodes. A Bayesian network represents a joint probability distribution factorised into component distributions. Figure 4 is an example of a simple Bayesian network.

Markov blankets appear when we consider the *conditional independence* of nodes with respect to one another. Conditional independence is a ternary relation, taking three relata[5]:

$$X \text{ is } \textit{conditionally independent} \text{ of } Z, \text{ given } Y, \text{ if for all } x, y, z : \; p(x|y, z) = p(x|y)$$

Intuitively, getting information about Z does not yield information about X if you already have information about Y. For example, let $p(x)$ be the probability of having an accident while driving, $p(z)$ the probability of wearing a coat, and $p(y)$ the probability of rain. Although there may be an increase in car accidents when the driver is wearing a coat, $p(x|z) > p(x)$, this can be explained by appealing to rain as a common cause of both. Knowing that it's raining means that you know there is an increased chance of an accident, but knowing *additionally* that the driver is wearing a coat does not further change this probability: $p(x|z, y) = p(x|y)$. Accidents are conditionally independent of coat-wearing, given rain.

[5] We use capital letters X, Y, Z to denote statistical variables and lower case letters x, y, z to denote the values of those variables. Conditional independence is a relation between variables, expressed as a certain equality holding for all values of those variables.

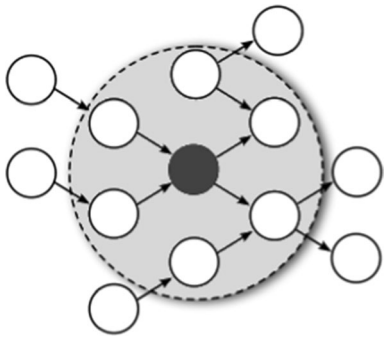

Fig. 5 A Markov blanket in a causal model. The Markov blanket of the grey node is the set of white nodes within the dotted circle. These are the parents (causal antecedents) and children (causal effects) of the grey node, as well as the parents of its children

A Markov blanket of a focal node X is simply a set of nodes such that X is conditionally independent of every other node in the graph with respect to that set. In Fig. 4, a Markov blanket of the focal node X_2 is the set of nodes $\{X_1, X_3, X_4\}$. Because of the way conditional independence is defined, knowing the values of the nodes in a Markov blanket means you will always know the value of its focal node—regardless of the value of any other node in the system.

Friston tries to augment the explanatory power of Markov blankets by discussing them in a causal rather than merely probabilistic setting (Friston 2013; Kirchhoff et al. 2018). Because causal models use the same underlying mathematical object as Bayesian networks—an *acyclic directed graph*—it is possible to carry over formally defined concepts from one setting to the other. In causal models, a Markov blanket for a particular node is the set of nodes that screen it from causal interactions with the rest of the system. As with probabilistic models, the values of the nodes in the blanket contain all the information required to know the state of the node in question (Fig. 5). Treating a causal model as a representation of a physical scenario, Friston interprets Markov blankets as physical boundaries separating the inside of a system from its outside. For example, one could treat a causal model as representing the interactions between a bacterium and its environment. Friston asserts that the cellular membrane is well-modelled by a Markov blanket whose focal node corresponds to the insides of the cell: "if we consider short-range electrochemical and nuclear forces, then a cell membrane forms a Markov blanket for internal intracellular states" (Friston 2013, p. 5). This modelling gambit embodies the claim that the bacterium's insides are conditionally independent of everything outside it, with respect to its membrane. This is an idealization, obtained by assuming that only short-range forces are relevant.

When a biological system is represented as a causal model, it appears to maintain a kind of stability within its Markov blanket. This can be contrasted with non-biological systems that tend to dissipate. Consider for example two systems suspended in water: a droplet of ink and a bacterium. The ink will rapidly diffuse. Globules initially inside its boundary will very soon come into contact with the water. Modelled causally, the Markov blanket of any node within the droplet will not remain intact for long: the boundary diffuses, and nodes previously inside the boundary will come into direct contact with the external environment. In contrast, the transaction

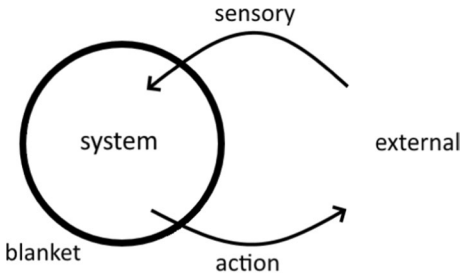

Fig. 6 A cyclical model of a Markov blanket system. A system with a Markov blanket is coupled to its environment via sensory input states and action output states

of materials and energy between the inside and outside of the bacterium will be more or less controlled by its cellular membrane. Part of the process of survival is maintaining such a membrane. Modelled causally, the bacterium's membrane is a Markov blanket.

Markov blankets, cyclical models, and phase spaces

To understand the conditions underlying the free energy principle, we first need to understand a related but different way of representing Markov blanket systems.

Discussions of the free energy principle do not usually employ causal models. Instead, they use diagrams depicting cyclical interactions between a system and its environment, as in Fig. 6. These are intended to emphasise the feedback inherent in perception-action loops. Causal models disallow cyclical relationships, because effects cannot be causally upstream from their causes. So Fig. 6 is not a causal model in the strict sense. Nonetheless, its relationship to causal models will be important later on.

Using the cyclical model in Fig. 6, biological systems can be decomposed in an important way. Continuing the bacterium example, the blanket separates the whole bacterium-water scenario into *internal* and *external* nodes. Internal nodes are those internal to the blanket. External nodes are those external to the blanket. We can furthermore divide the blanket itself into two kinds of node: *sensory* and *active*. Sensory nodes are affected by external nodes and only affect internal and active nodes. Active nodes are any that affect external nodes (even if they also have internal effects). From these definitions, internal nodes are causally segregated from external nodes by the blanket.

Let us define the *state* of a system as the collection of values of its nodes. For example, if each of the sensory, internal and action nodes can take one of two values, then the overall state of a three-node system is defined as a list of three values:

$$\text{state} = \langle \text{sensory node, internal node, action node} \rangle$$

and there are eight possible states, corresponding to the eight combinations of values the nodes can take: $\langle 0, 0, 0 \rangle$, $\langle 0, 0, 1 \rangle$, $\langle 0, 1, 0 \rangle$ and so on.

Over time, as the value of each component node changes, the system as a whole changes, moving from state to state. It changes as a result both of the influence of the external environment on it, and its own actions. Its states—all the values of its nodes

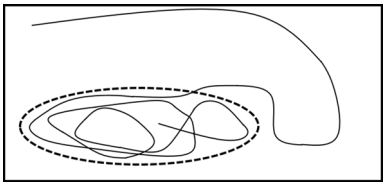

Fig. 7 A phase space diagram depicting a system's trajectory. The entire rectangle represents the phase space of the system, a multidimensional space each point of which denotes a distinct state the system could be in. The solid curve represents the system's trajectory over time. When it enters the region defined by the dashed ellipse, it nevers leaves, instead moving around this region. The region is a global attractor (The trajectory crosses over itself which would be strictly impossible in a deterministic setting; we can imagine a third dimension that the system is moving through)

at any given time—define a *phase space*. Each point of a phase space completely defines the state of a system at a given time. For three nodes each with two possible values, the phase space has eight points, which could be envisioned as the vertices of a cube. A more realistic model would contain vastly more possible points.

Depending on the dynamics of the entire scenario—blanketed system plus environment—the system may visit some of these points more often than others. Suppose we have a cyclical model whose nodes change over time according to a dynamical rule. Suppose also that there is a system within the model that possesses a Markov blanket. The system's nodes constitute a phase space, which we could plot, and trace its trajectory (Fig. 7).

Equivalence classes relate the three model types

Drawing a diagram like Fig. 7, and plotting the trajectory of a system, requires that it be possible for the system to revisit the same states at different times. This corresponds to the cyclical model in Fig. 6 being 'run forward' in time according to a dynamic rule, its node values being plotted, and it sometimes having the same node values at different points in time. Notice that this kind of situation is not strictly possible in a *causal* model. Bayesian networks and causal models are *acyclic*: if you trace a path along the graph following the direction of the arrows, you can never get back to where you started. The 'flow of time' in a causal model is one way; setting nodes to certain values affects downstream nodes but not those upstream. As a result, causal models do not inherently support the notion of identity over time. A node corresponding to a component of a system at time t need not represent the same part of the system at one of its child nodes at time $t + 1$. Given that we are going to represent proper functions using causal models, and the free energy principle using dynamic cyclical models, this presents a challenge to our attempt to relate the two sets of conditions.

Fortunately, there is a way to relate causal models to cyclical models, and thus to the kind of phase space diagram in Fig. 7. Each node in the cyclical diagram can be mapped to a causally linked chain of time-indexed nodes in an acyclic causal graph (Fig. 8). In other words, the cyclic model can be 'unrolled' into an acyclic time-indexed causal graph, which can then be treated as a Bayesian network or causal

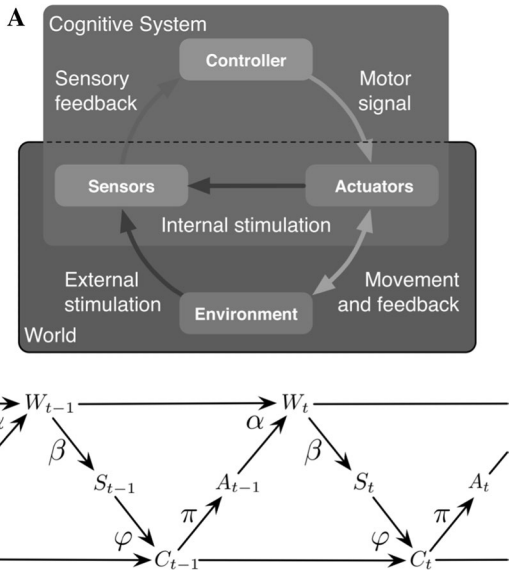

Fig. 8 The relationship between cyclical models (**A**) and causal models (**B**). A cyclical model can be converted into a causal model by 'unrolling' it through time, treating the same node as different at each timestep. Conversely, a causal model can be converted into a cyclical model by identifying equivalence classes of nodes, those that represent the same part of the system at different timesteps. Adapted from figures 3 and 4 of Ay and Zahedi (2014, p. 266). *W* environment, *S* sensors, *C* controller (internal nodes), *A* actuators, *t* time index, α, β, π, ϕ causal effects determined by the dynamical rules governing the system

model. Sometimes more complex diagrams used in the active inference literature explicitly adopt this 'unrolled' form (see e.g. Friston et al. 2017, Fig. 2, p. 389). In the opposite direction, a causal model can be translated into a cyclical model by identifying *equivalence classes*: sets of nodes to be treated as representing the same component of the system at different time stages. Identifying equivalent nodes 'rolls up' the system, producing the cyclical form apparent in Fig. 6.

Furthermore, phase spaces are produced from cyclical models by a similar process of identifying equivalence classes. This time, each equivalence class is defined in terms of the system being in the same state at different times. Phase space diagrams treat all states as identical, plotting them at the same point in the space, regardless of the time at which the system reaches that state. In the other direction, converting a phase space diagram into a cyclical model requires that a trajectory be drawn in the space. This trajectory can then be treated as representing the flow of time in the corresponding cyclical model. Figure 9 depicts these formal relationships between the three kinds of model.

Conditions under which the free energy principle holds

We do not need to state the principle itself at this stage, just the conditions under which it holds. Consider again our blanketed system depicted in the cyclical model (Fig. 6). Because the system's states change both as a result of its actions and external affairs, its trajectory can be explained by reference to both of these things. In particular, aspects or patterns in its trajectory could be explained by reference to the system's actions. That is the key idea behind the free energy principle.

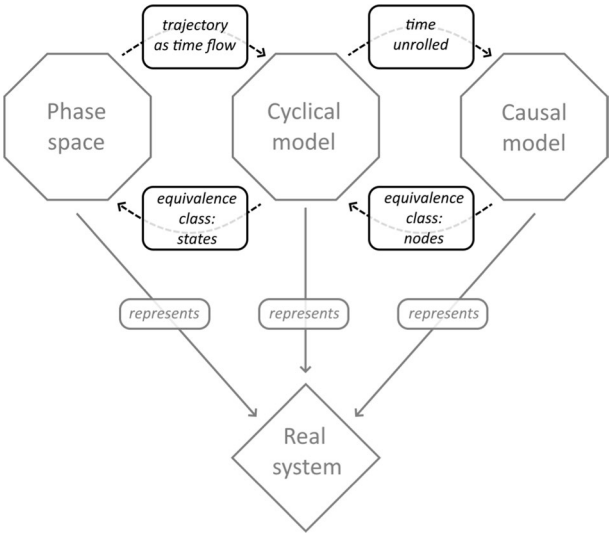

Fig. 9 Three kinds of model can be transformed into each other. *Moving from left to right*: A trajectory within a phase space diagram defines the time flow of a dynamic, cyclical model. The cyclical model can then be 'time-unrolled' to produce a causal model (see Fig. 8). *Moving from right to left*: Within a causal model, equivalence classes of nodes can be defined, such that treating those nodes as identical 're-rolls' the model, allowing the system in question to undergo change over time. Finally, a system in a dynamic, cyclical model—which adopts different states at different times—defines a phase space diagram when equivalent states are assigned to the same point in the space; the system's change over time becomes a trajectory in the phase space

Suppose our system eventually ends up in a kind of fluctuating cycle, inhabiting a distinct set of states that it wanders through but never leaves (Fig. 7, dashed ellipse). Friston calls this cycle an "invariant set of states" or a "random global attractor" (Friston 2013, p. 3); later it is described as "non-equilibrium steady-state" (Da Costa et al. 2021, p. 3). When a system with a Markov blanket converges to an attractor, the free energy principle holds for that system.

At first glance it looks as though the lack of identity over time in causal models might prevent these conditions being defined. An anonymous reviewer suggested to us that since the time-unrolled system does not have downstream nodes that are identified with upstream nodes, there is no sense in which the system can 'revisit' states. There are lots of time slices of a system that look similar, but cannot be identified (within that causal model) as the same system. And because there is no single system moving through different states, a phase space diagram cannot be drawn for it; there can be no probability distribution over the different states it occupies, and so there is no way to define a non-equilibrium steady-state for it.

It seems to us that this issue is best resolved by emphasising the relations between the three kinds of model. Causal models, with time-unrolled collections of nodes, are useful for discovering causal effects and pathways. Defining a non-equilibrium steady-state is achieved by rolling the model, first into a cyclical model by identifying equivalence classes of nodes, then into a phase space diagram by identifying equivalent states. Conditions on the free energy principle are definable via cyclical

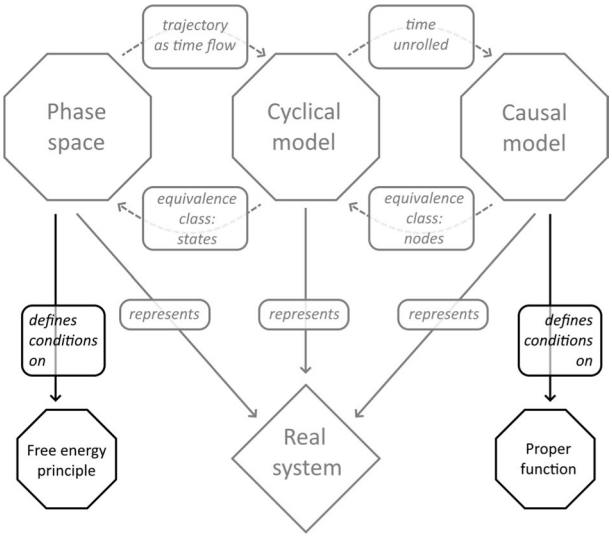

Fig. 10 The free energy principle and proper functions are formally defined in terms of different kinds of model

models and phase spaces, while proper functions (to which we shall soon turn) are definable in terms of causal models (Fig. 10). Manipulating our formal depiction of a system in order to derive three different descriptions is not inconsistent. It simply reveals that the system can be understood in different, yet compatible, ways.

Later we will consider the content of the free energy principle, which describes the consequences when the conditions just introduced hold. What matters here is that a system can remain within an attractor, and part of what explains why the system remains in the attractor is the system's own active states: how it acts on the external world. This is where a connection with proper functions is found.

Conditions on possession of a proper function

To define proper function we need to introduce three prior definitions, all given by Millikan (1984). First is a definition of *reproduction*:

An individual B is a "reproduction" of an individual A iff:

1. B has some determinate properties p_1, p_2, p_3 etc., in common with A;
2. That A and B have the properties p_1, p_2, p_3 etc., in common can be explained by a natural law or laws operative in situ;
3. For each property p_1, p_2, p_3 etc., the laws in situ that explain why B is like A in respect to p are laws that correlate a specifiable range of determinates under a determinable under which p falls, such that whatever determinate

> characterizes *A* must also characterize *B*, the direction of causality being straight from *A* to *B*.
>
> Excerpted from Millikan (1984, pp. 19–20)

The final condition is a little hard to parse. Intuitively it says that the reason *B* has the properties it does is that *A* has the properties it does. Millikan paraphrases,

> Roughly, the law in situ implies that *had A been different* with respect to its determinate character *p* within a specifiable range of variation, as a result, *B would have differed accordingly*.
>
> Millikan (1984, p. 20), emphasis original

Although the definition distinguishes individuals *A* and *B*, an offspring-parent relationship is not strictly required. We propose to let different time-stages of the same system, represented by different nodes in a causal model, count as different individuals for the purposes of the definition. As a result, unrolling the time-stages of Friston's cycle within its attractor produces a causal model in which causally linked 'copies' of time-stages count as reproductions of each other. When the system returns to a point in phase space, its current time-stage is a "reproduction" of its time-stage when it was last at that point.[6] To see that time-stages are reproductions as defined above, consider each part of the definition in turn. Condition (1) is satisfied because *B* is defined as the system at the same point in the phase space as *A*. One of the things guaranteed by the relation "being at the same point in the phase space as" is having properties in common. Indeed, a system's position in a phase space *defines* property values of the system at that time. Condition (2) is satisfied because the attractor set is determined by the structure of the system and the laws governing its time evolution.[7] Furthermore, attractor cycles define equivalence classes of system states. If system state *A* had differed, everything in its equivalence class would have differed in the same way. So condition (3) holds, and later time-stages at the same point in state space are reproductions of earlier ones.

The next two preliminary definitions can be given briefly. Again we take Millikan's strict definitions and interpret them in the context of a causal model. First is *reproductively established family*:

> Any set of entities having the same or similar reproductively established characters derived by repetitive reproductions from the same character of the same model or models form a *first-order reproductively established family*.

[6] An anonymous reviewer complained (rightly) that our interpretation in this section stretches Millikan's notion of proper function almost to breaking point. Indeed the insistence on reproduction has led to dissatisfaction with Millikan's strict account and has motivated extended definitions of functions underpinning teleosemantics, most notably by Shea (2018, §3). We are stubbornly cleaving to Millikan's definition, but we leave open the possibility that alternative definitions do the job better. These issues belong also to a wider debate in the philosophy of biology, on which we really do not have space to comment, about selected effects, reproduction, and persistence (Bouchard 2014; Bourrat 2021).

[7] For example, in Friston (2013) the laws are encapsulated by a differential equation called the Fokker–Planck equation. In Da Costa et al. (2021) the laws are defined in terms of a stochastic process called the Ornstein-Uhlenbeck process.

> Millikan (1984, p. 23), emphasis original

It should be clear that equivalence classes define groups of nodes that play the role of reproductively established families. We might imagine Friston's system spiralling into its attractor. Each time it passes through a given region of phase space, it belongs to the equivalence class defined in terms of the properties that characterise that region. The bacterium time-stage at time t in state A—say, with node values $\langle 1, 0, 1\rangle$—belongs to an equivalence class with any other time-stage, both before and after t, that is also in state A.

One more preliminary definition is that of an *ancestor*:

> Any member of a (first-order) reproductively established family from which a current member m was derived by reproduction or by successive reproductions is an ancestor of m.
> Millikan (1984, p. 27)

Every earlier time-stage of the system within the current time-stage's equivalence class counts as an ancestor.

When the above definitions apply, phase-space diagram systems with Markov blankets that converge to an attractor correspond to causal-model systems that have proper functions. To see this, consider the full definition:

> Where m is a member of a reproductively established family R and R has the reproductively established [...] character C, m has the function F as a [...] proper function iff:
>
> 1. Certain ancestors of m performed F.
> 2. In part because there existed a direct causal connection between having the character C and performance of the function F in the case of these ancestors of m, C correlated positively with F over a certain set of items S which included these ancestors and other things not having C.
> 3. One among the legitimate explanations that can be given of the fact that m exists makes reference to the fact that C correlated positively with F over S, either directly causing reproduction of m or explaining why R was proliferated and hence why m exists.
>
> Millikan (1984, p. 28)

Before stepping through each of the three conditions, some comments are needed. First, a system's reproductively established character C is the group of properties that define reproduction. They are the properties that characterise each equivalence class of time-stages. Second, F is a causal effect. That there is a causal link between C and F explains proliferation of the family having character C, hence continuing occurrence of F.

Finally, the set S is a contrast class. Persistence of the family can be explained by reference to S. To identify a set to play this explanatory role, we need to consider other systems that did not find themselves in attractors. Our earlier example of an ink blot in water contrasts with successfully converging systems, but

we are focusing only on models rather than real systems in order to sidestep the question of realism that Friston must eventually face up to. For now, we can just say that the contrast class consists of models that do not include systems with Markov blankets that persist over time. After all, a satisfactory explanation of why a system *with* a Markov blanket persists over time should entail that systems without one do not persist in the same way.

Friston's systems satisfy Millikan's conditions

Now we can show that our Markov blanket system within its attractor possesses a proper function. We will show that the boundary itself has a proper function to effect certain causal changes in the external state. Let m be the current time-stage of the whole blanket system; C the system's boundary at the current time-stage; R the family of previous boundary time-stages in the same equivalence class:

1. Certain ancestors of m performed F.
   - YES: let F be the causal effects of the blanket system on the external system. By assumption the system has these effects because the Markov blanket can be divided into sensory and active nodes. The relevant causal effects are those of the active nodes on external nodes.
2. There is a causal connection $C \rightarrow F$ and for this reason C correlated positively with F over a set S, where S includes ancestors of m as well as things that didn't have C.
   - YES, because causal effects of the internal nodes on external nodes must go through the boundary, so F must be affected by C. Our system ended up in the attractor when other systems did not. We said above that the system's actions are part of what explains why it remains in the attractor; therefore, its actions (constituting the causal connection $C \rightarrow F$) are part of the reason why the system remained within the attractor (why C correlated with F over set S).
3. Explanation of the existence of m can refer to the positive correlation between C and F.
   - YES, because the boundary is a causal structure doing real explanatory work on Friston's account.

The variables C and F will be filled in differently for each system. But they must be filled in for the free energy principle to hold. Once they are filled in, F is a proper function of m. There are external causal effects of the boundary that are proper functions of the blanketed system. Figure 11 extends Fig. 10 to depict the equivalence of these conditions as a consequence of the relationship between causal models, cyclical models, and phase space diagrams.

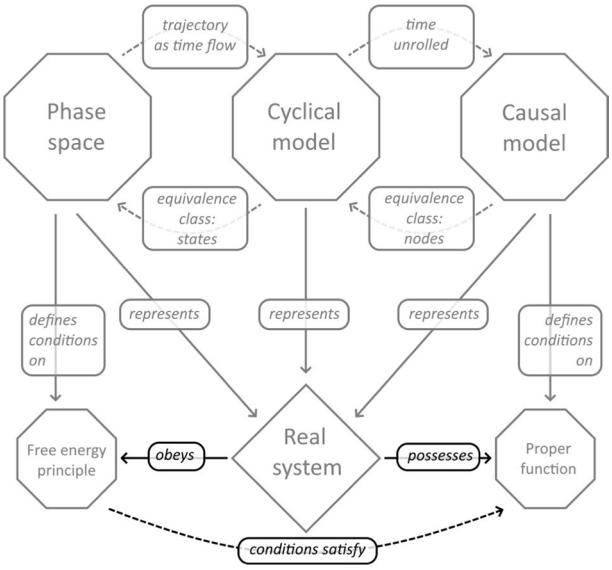

Fig. 11 Conditions on obeying the free energy principle satisfy conditions on possession of a proper function. Because phase space diagrams and causal models can be related to each other via equivalence classes, the conditions on a phase-diagram system obeying the free energy principle can be shown to satisfy the conditions on a related causal model system possessing a proper function. Real systems faithfully depicted by these models would satisfy both sets of conditions. We have not investigated the converse satisfaction

Are functions of time-stages functions of whole systems?

An anonymous reviewer raised the following problem. We have shown only that different time-stages m_t of a system have different functions F_t. We have not shown that the system m has each of these functions at different times, nor even that there is some aggregate function F that the system has. It is not as though we have posited a function of a subsystem—say, a face-recognizer in the brain—and then posited that same function as belonging to the wider system—the brain. Furthermore, even if we are right to claim that time-stages count as ancestors of each other, the system as a whole cannot have its own time-stages as ancestors. It may have ancestors, and thereby a selection history, but that would be a different history to the one we have described.

To respond, we think that if progressive time-stages can be identified as belonging to a unified system, then that system ought to be attributed the functions attributed to the time-stages. We agree that the situation is unlike positing that the function of a subsystem also belongs to the supersystem, like in the face-recognizer case. But we do not think our point rests on such an analogy. Rather, the relationship between a system and its time stages is even more intimate than that between a system and its components. It is difficult to see how we could

avoid attributing time-stage functions to the system as a whole. What is doing the work here is the assumption, implicit in the association of the three models with each other, that the time-stages do indeed belong to the same system.[8]

Whether or not a system that has different functions F_t at different times thereby has some unified function F is a separate issue. It depends how reasonable it is to unify all those disparate performances under a single description. This is a question that has been discussed in the wider literature on teleosemantics, and we do not propose to answer it here (Millikan 1990). We can instead try to make the required result plausible with an example. Consider a system that needs to avoid danger (perhaps a predator or ambient toxin) and can sense one of two conditions: either there is danger to the north or to the south. When there is danger to the north it must move south, and vice versa. Suppose that at $t = 1$ there is danger to the north. Then a function of its time-stage at F_1 is to move south. Now suppose at $t = 2$ there is danger to the south. Then a function of its time-stage at F_2 is to move north. These are different functions. We suggest that it is not unreasonable to posit a more general function, both to the time-stages and to the system as a whole, to *move away from danger*. We posit further that the kinds of examples of functions that could be employed to illustrate our account will all be of this nature. The alternative would be for a system to have entirely distinct functions at different times, none of which could be generalised under a common description. In an environment that demanded such disparate activities of a single entity, it is unlikely anything could survive long enough to satisfy the conditions on the free energy principle or possession of a proper function.

It is time to move on. In the next section we describe how Friston's and Millikan's formulations give rise to interesting relations between the internal and external states of a system.

Signals as internal models

In this section, we suggest a correspondence between Ruth Millikan's teleosemantic definition of *signal* and Karl Friston's purported proof that biological entities *will appear to model* the external world.

First consider implicit modelling on Friston's account. Friston claims to be able to show that the internal state of a blanketed system must appear to model the external state (Friston 2013; Da Costa et al. 2021). What does he mean? A physical system implicitly models another physical system when the state of the first *parametrizes a*

[8] This issue is related to the problem of function attribution for merely persisting (i.e. not reproducing) systems; see footnote 6. What we require is that *differential persistence due to causal effects* is sufficient for bestowing a proper function. If that requires extending the definition, so be it. The explanation of why the system continues to exist appeals to its past actions and contrasts them with systems that did not act this way, and so did not persist. That pattern of explanation is what is driving both the free energy principle and the definition of proper functions. In the worst case, our argument can be read as a conditional: *if* there is a workable notion of etiological function for merely persisting systems, *then* conditions on the free energy principle satisfy conditions on possession of that kind of function.

distribution over states of the second. This means that the different possible states of the first system (the one doing the modelling) can be mapped onto mathematical values which determine a probability distribution over the different possible states of the second system (the one being modelled). Any two statistically correlated systems could be said to parametrize distributions over each other in this sense. In the example in "Markov blankets" section, the collection of drivers wearing coats could be said to implicitly model locations of car crashes. We just need to define a function that maps from the geographical locations of coat-wearers to a probability distribution over crashes in those locations. This is not a consequence of any inferential or even causal relationship between the two (although in this case they have a common cause), just a consequence of their statistical relationship.

In the case at hand, the internal state corresponds to a mathematical object which expresses the probability that the external state is a certain way. The idea is that implicit modelling, via this kind of parametrization, is assigned a stronger explanatory role than mere statistical correlation for blanketed systems. Friston and colleagues appeal to this implicit model to explain behaviour (Da Costa et al. 2021, § 4). In order to persist within the attractor, the blanket system must match its implicit model to the world. So changes in internal states can be explained as the system's attempt to accurately model the external state. Matching the model to the world helps the blanket system survive. The implicit model is variously described as "a probabilistic representation (recognition density) encoded by the agent..." (Friston 2009, p. 294), "a probability density over external states [...] that is encoded (parametrized) by internal states" (Friston 2013, p. 4), and "...probabilistic beliefs that are implicit in a system's interactions with its local surroundings" (Kirchhoff et al. 2018, p. 2).

Now we can state the free energy principle. When the conditions described in "Friston's systems satisfy Millikan's conditions" section hold, blanket systems will "appear to minimize free energy" (Friston 2013, p. 2). Here "free energy" has nothing to do with energy in the traditional, physical sense. It is a statistical property rather than a physical one. Free energy describes the unlikelihood of the sensory states that the system is currently receiving, where the probability distribution that defines this unlikelihood is constructed from the system's historical trajectory. The free energy principle says that systems that persist out of thermodynamic equilibrium will appear to minimise this unlikelihood. Intuitively, surviving systems are those that inhabit environments similar to those inhabited by their successful ancestors. It is as a result of this minimization that the implicit model will come to match the world.

With a little more work we can interpret Millikan's definition of 'signal' such that there exist representational relations between internal states of the blanket system and the world.[9] Above we saw that representational relations come in two kinds: *descriptive* and *directive*. We suggest that it is possible to treat the internal state as

[9] Millikan calls signals "intentional icons" (Millikan 1984, p. 96ff), and later "representations" (Millikan 2004, p. 77ff).

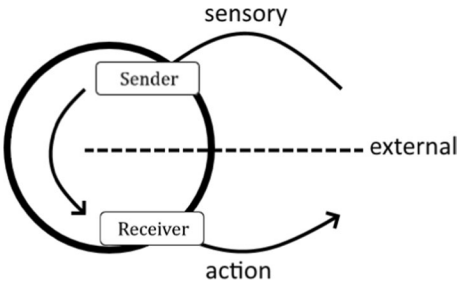

Fig. 12 A teleosemantic interpretation of the free energy principle. A Markov blanket system induces sender–receiver structure, suggesting a teleosemantic analysis. Friston claims that the inner state will implicitly model the external state. Teleosemantics says the inner state, treated as a signal, bears a representational relation to the external state

a signal, and that its descriptive representational relation is comparable to Friston's notion of implicit modelling.

On Millikan's account, a signal is an intermediary between a coadapted sender and receiver. The representational relations borne by the signal to its content can be analyzed in terms of its causal effects. The directive content is the effect that should be brought about by the receiver (in Fig. 2, this is the Effect variable). The descriptive content—what the signal is usually said to represent—is what would have to be the case such that the receiver successfully performs its proper function (in Fig. 2, this is the State variable). In other words, the representational content is whatever external circumstance must be the case in order for the effects of action to be successful in accordance with proper function.

We need to match the sender–receiver framework to our blanket system. This can be done by casting the sender as the sensory state, the signal as the inner state, and the receiver as the active state (Fig. 12). In our blanket system, sender, signal and receiver are clearly coadapted. They are bound within the same Markov blanket that has reached an attractor. They have a unified proper function. Although they may not be materially distinct, Friston gives a functional analysis that we are assuming individuates subsystems by their proper functions. The division of labour between sensory, internal and action states is required both for Friston's mathematical proofs and Millikan's definition.

Continuing with Millikan's definition, consider the directive mapping relation. Suppose the inner state, as affected by the senses, takes some physical form that causes the action state to behave in a certain way. The action produces an effect that helps keep the blanket system in the attractor. That is, it performs one of its proper functions. The external changes are dependent on the form the inner state took. So there is a relationship between inner state and external effect, mediated by the active state. This is the directive aspect of the inner state:

> *Directive aspect*: the external change caused by the system's behaviour that constitutes its successfully performing its proper function.

Persistence of the system is explained, in part, by a directive mapping relation that relates various different inner states to various different external effects that action is supposed to produce.

Consider now the more familiar descriptive representational aspect of the inner state. It has to be characterized in a little more detail. It cannot simply be what caused the inner state. Instead:

> *Descriptive aspect*: the external state that must obtain for the effects of action to successfully perform the system's proper function.

In simple cases this might well be equivalent to what caused the inner state.

The signal corresponds to what must be the case for action to successfully produce the required outcome. This suggests that differences in the inner state correspond to differences in the external state: differences in inner state lead to different actions, which will be successful if the external state is appropriate (i.e. promotes the proliferation of the blanket system) given that action. And one way to describe differences in inner state corresponding to differences in external state is to say that the inner state is an implicit model of the external state.

On both Millikan's and Friston's accounts, there is a system whose survival-promoting behaviours can be in part explained in terms of the system harbouring an implicit model of its surroundings. Friston describes this model in terms of the "recognition density" (2009, p. 293) or "probability density over external states" (2013, p. 4). Millikan describes the implicit model as the representational content of the inner state considered as a signal. We believe that this concordance between their accounts is surprising, and worthy of further study.

A couple of comments are in order before closing. In active inference, there are two different kinds of activity: action, which minimizes expected free energy, and inference, which minimizes variational free energy. We suspect this corresponds to matching the world to the model (acting so the receiver's proper function is satisfied) and matching the model to the world (updating the signal so its representational content is correct). A similar claim is made by Hohwy (2013, §4) in the context of predictive processing theories of cognition. In simpler, tightly coupled systems, these two kinds of behaviour are barely distinguishable (Fig. 12).

The big difference between the free-energy and proper-functional descriptions of a system is the answer given to the question 'what is the system trying to do?' Friston says all systems are minimizing the free energy of their sensory states. Millikan says all systems have different proper functions, but at the most abstract level a system's ultimate proper function is not to minimize free energy but to reproduce. From one perspective, reproduction is a means to minimizing free energy. From the other, minimizing free energy is a means to reproduction.

Future research and closing remarks

Our analysis points to various lines of further research. For example, there is an active debate concerning whether or not the free energy principle should be given a representational reading (Gładziejewski 2016; Gładziejewski and Miłkowski 2017;

Kiefer and Hohwy 2018), or if it should be understood in enactivist (anti-representational) terms (Ramstead et al. 2021; Gallagher and Allen 2018; Sims and Pezzulo 2021). The congruence with teleosemantics here suggests a representational reading. Indeed, for proponents of teleosemantics, it might *force* a representational reading. Another line of future research concerns the issue of reduction. Does the free energy principle *reduce to* proper functions, or vice versa? A third observation concerns the ever-widening scope of the free energy principle. If the framework encompasses domains like culture (Rubin et al. 2020; Veissière et al. 2020), must these domains likewise have proper functions?

There is not the scope here to engage in detail with the implications of the formal links described in this paper. However, we do want to finish with a brief discussion of one of the most prominent conceptual problems levelled at teleosemantic theories, and explore how our analysis might inform a response to this problem.[10] Broadly, the challenge concerns whether past processes can account for representational content in the here-and-now. This is typically illustrated by the "Swampman" thought experiment. In this case, a random lightning strike in a swamp produces an intrinsic duplicate of a human. Our intuitions, it is claimed, indicate that Swampman would act exactly like a person with contentful representational states. But as teleosemantic theories rely on history, and Swampman has no history, we cannot explain his behaviour by recourse to mental content. This result is thought to push us toward internalist theories of function, and hence content.[11] Does our analysis provide support for the teleosemantic view?

Teleosemantics says that at the immediate point of creation Swampman possesses neither proper functions nor representational content. So much is familiar. Our account emphasises a further point we have not yet seen in the literature: at the point of creation, Swampman is not subject to the conditions of the FEP. Swampman cannot be said to minimise free energy, because he has no historical trajectory from which to construct the probability distributions that define free energy in the first place. As Swampman moves around in the world, however, he starts to conform to the conditions of the free energy principle: by reliably persisting within his Markov blanket, historical tallies of sensory states define the requisite probability distributions. Over time, it becomes possible to define Swampman's free energy, and to explain his continuing survival by reference to his minimising that quantity. As we have seen, any system that meets the conditions of the FEP satisfies the conditions

[10] Thanks to an anonymous reviewer for pushing us on this issue.

[11] Here we are engaging with the Swampman challenge at a first-order level. But we take very seriously second-order, methodological challenges concerning the viability of modally immodest thought experiments. First, there are well-known challenges to this philosophical approach (Machery 2017). Second, whether or not you take Swampman seriously will depend on how important you think it is for a philosophical theory to capture all our folk intuitions. If, for instance, we put more weight on the practical upshots of our theory, then the fact that it fails to account for our intuitions in some modally extreme cases will not matter very much (Woodward 2021, pp. 28–35). In general, we agree with this more pragmatic approach. We engage with the thought experiment because, as will become clear, we think it nicely illustrates the conceptual links between teleosemantics and the free energy principle that we have identified in the paper.

on possession of a proper function, so this entails that Swampman will slowly start to gain proper functions.

As it happens, at least some teleosemanticists have endorsed this result. Nick Shea's account of content, based on a kind of etiological function called *task functions*, has the same consequence:

> As soon as a swamp system starts interacting with its environment and learning, it will rapidly acquire task functions. So, it won't be long before there is a basis for counting some outcomes as successful and others as unsuccessful, and then we can start explaining the success and failure of its behaviour in terms of correct and incorrect representation.
> Shea (2018, p. 169)

There is clearly much more to be said on these topics. However, this brief discussion indicates how the correspondence between the FEP and teleosemantics can be used to illuminate one of the key problems in the literature on historical functions. We hope further analyses will be forthcoming.

Acknowledgements Thanks to Kim Sterelny and Michael Kirchhoff for comments on earlier drafts, and two anonymous reviewers for comments on subsequent versions. SFM would like to thank Russell Gray and Iren Hartmann for generously organising a guest researcher position at the Max Planck Institute for Evolutionary Anthropology, during which time much of the manuscript was written.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors declared that they have no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Artiga M (2010) Learning and selection processes. THEORIA 25(2):197–209

Ay N, Zahedi K (2014) On the causal structure of the sensorimotor loop. In: Prokopenko M (ed) Guided self-organization: inception. Springer, Berlin, pp 261–294

Baigrie BS (1989) Natural selection vs trial and error elimination. Int Stud Philos Sci 3(2):157–172

Bouchard F (2014) Ecosystem evolution is about variation and persistence, not populations and reproduction. Biol Theory 9(4):382–391

Bourrat P (2021) Function, persistence, and selection: generalizing the selected-effect account of function adequately. Stud Hist Philos Sci Part A 90:61–67

Bruineberg J et al (2021) The emperor's New Markov blankets. Behav Brain Sci. https://doi.org/10.1017/S0140525X21002351

Catania A (1999) Thorndike's legacy: learning, selection, and the law of effect. J Exp Anal Behav 72(3):425–428

Colombo M, Cory W (2018) First principles in the life sciences: the free-energy principle, organicism, and mechanism. Synthese 198:3463–3488

Da Costa L et al (2021) Bayesian mechanics for stationary processes. arXiv Preprint. arXiv:2106.13830 [math-ph, physics:nlin, q-bio]

Friston K (2009) The free-energy principle: a rough guide to the brain? Trends Cogn Sci 13(7):293–301

Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11(2):127–138

Friston KJ (2013) Life as we know it. J R Soc Interface 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

Gallagher S, Allen M (2018) Active inference, enactivism and the hermeneutics of social cognition. Synthese 195(6):2627–2648

Gładziejewski P (2016) Predictive coding and representationalism. Synthese 193(2):559–582

Gładziejewski P, Miłkowski M (2017) Structural representations: causally relevant and different from detectors. Biol Philos 32(3):337–355

Hohwy J (2013) The predictive mind. Oxford University Press, Oxford

Hohwy J (2020) Self-supervision, normativity and the free energy principle. Synthese 199(1–2):1–25

Hull DL, Langman RE, Glenn SS (2001) A general account of selection: biology, immunology, and behavior. Behav Brain Sci 24(3):511–573

Kiefer A, Hohwy J (2018) Content and misrepresentation in hierarchical generative models. Synthese 195(6):2387–2415

Kingsbury J (2008) Learning and selection. Biol Philos 23(4):493–507

Kirchhoff MD (2018) Autopoiesis, free energy, and the life-mind continuity thesis. Synthese 195(6):2519–2540

Kirchhoff M et al (2018) The Markov blankets of life: autonomy, active inference and the free energy principle. J R Soc Interface 15(138):20170792

Levins R (1966) The strategy of model building in population biology. Am Sci 54(4):421–431

Machery E (2017) Philosophy within its proper bounds. Oxford University Press, Oxford

Millikan RG (1984). Language, thought, and other biological categories. MIT, Cambridge

Millikan RG (1989) Biosemantics. J Philos 86(6):281–297

Millikan RG (1990) Truth rules, hoverflies, and the Kripke–Wittgenstein paradox. Philos Rev 99(3):323–353

Millikan RG (2004) Varieties of meaning. MIT, Cambridge

Morowitz H (1986) Entropy and nonsense. Biol Philos 1(4):473–476

Neander K (2017) A mark of the mental: in defense of informational teleosemantics. MIT, Cambridge

Ramstead MJD et al (2021) Neural and phenotypic representation under the free-energy principle. Neurosci Biobehav Rev 120:109–122

Rubin S et al (2020) Future climates: Markov blankets and active inference in the biosphere. J R Soc Interface 17(172):20200503

Shea N (2018) Representation in cognitive science. Oxford University Press, Oxford

Sims M, Pezzulo (2021) Modelling ourselves: what the free energy principle reveals about our implicit notions of representation. Synthese 199:7801–7833

Skinner BF (1981) Selection by consequences. Science 213(4507):501–504

Sprevak M (2020) Two kinds of information processing in cognition. Rev Philos Psychol 11:591–611

Veissière SPL et al (2020) Thinking through other minds: a variational approach to cognition and culture. Behav Brain Sci 43:e90
Watson RA, Szathmáry E (2016) How can evolution learn? Trends Ecol Evol 31(2):147–157
Weisberg M (2006) Forty years of 'the strategy': Levins on model building and idealization. Biol Philos 21(5):623–645
Williams D (2021) Is the brain an organ for free energy minimisation? Philos Stud. https://doi.org/10.1007/s11098-021-01722-0
Woodward J (2021) Causation with a human face: normative theory and descriptive psychology. Oxford University Press, Oxford