INTERFACE FOCUS

rsfs.royalsocietypublishing.org

Research

Cite this article: Kolchinsky A, Wolpert DH. 2018 Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface Focus* **8**: 20180041. http://dx.doi.org/10.1098/rsfs.2018.0041

Accepted: 4 September 2018

One contribution of 10 to a theme issue 'Computation by natural systems'.

Subject Areas:
biocomplexity

Keywords:
information theory, semantic information, agency, autonomy, non-equilibrium, entropy

Author for correspondence:
Artemy Kolchinsky
e-mail: artemyk@gmail.com

THE ROYAL SOCIETY PUBLISHING

Semantic information, autonomous agency and non-equilibrium statistical physics

Artemy Kolchinsky[1] and David H. Wolpert[1,2,3]

[1]Santa Fe Institute, Santa Fe, NM 87501, USA
[2]Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Arizona State University, Tempe, AZ, USA

AK, 0000-0002-3518-9208

Shannon information theory provides various measures of so-called syntactic information, which reflect the amount of statistical correlation between systems. By contrast, the concept of 'semantic information' refers to those correlations which carry significance or 'meaning' for a given system. Semantic information plays an important role in many fields, including biology, cognitive science and philosophy, and there has been a long-standing interest in formulating a broadly applicable and formal theory of semantic information. In this paper, we introduce such a theory. We define semantic information as the syntactic information that a physical system has about its environment which is causally necessary for the system to maintain its own existence. 'Causal necessity' is defined in terms of counter-factual interventions which scramble correlations between the system and its environment, while 'maintaining existence' is defined in terms of the system's ability to keep itself in a low entropy state. We also use recent results in non-equilibrium statistical physics to analyse semantic information from a thermodynamic point of view. Our framework is grounded in the intrinsic dynamics of a system coupled to an environment, and is applicable to any physical system, living or otherwise. It leads to formal definitions of several concepts that have been intuitively understood to be related to semantic information, including 'value of information', 'semantic content' and 'agency'.

1. Introduction

The concept of *semantic information* refers to information which is in some sense meaningful for a system, rather than merely correlational. It plays an important role in many fields, including biology [1–9], cognitive science [10–14], artificial intelligence [15–17], information theory [18–21] and philosophy [22–24].[1] Given the ubiquity of this concept, an important question is whether it can be defined in a formal and broadly applicable manner. Such a definition could be used to analyse and clarify issues concerning semantic information in a variety of fields, and possibly to uncover novel connections between those fields. A second, related question is whether one can construct a formal definition of semantic information that applies not only to living beings but also *any* physical system—whether a rock, a hurricane or a cell. A formal definition which can be applied to the full range of physical systems may provide novel insights into how living and non-living systems are related.

The main contribution of this paper is a definition of semantic information that positively answers both of these questions, following ideas publicly presented at the FQXi's 5th International Conference [31] and explored by Carlo Rovelli [32]. In a nutshell, we define *semantic information* as 'the information that a physical system has about its environment that is causally necessary for the system to maintain its own existence over time'. Our definition is grounded in the intrinsic dynamics of a system and its environment, and, as we will show, it formalizes existing intuitions while leveraging ideas from information theory and non-equilibrium statistical physics [33,34]. It also leads to a

non-negative decomposition of information measures into 'meaningful bits' and 'meaningless bits', and provides a coherent quantitative framework for expressing a constellation of concepts related to 'semantic information', such as 'value of information', 'semantic content' and 'agency'.

1.1. Background

Historically, semantic information has been contrasted with *syntactic information*, which quantifies various kinds of statistical correlation between two systems, with no consideration of what such correlations 'mean'. Syntactic information is usually studied using Shannon's well-known information theory and its extensions [35,36], which provide measures that quantify how much knowledge of the state of one system reduces statistical uncertainty about the state of the other system, possibly at a different point in time. When introducing his information theory, Shannon focused on the engineering problem of accurately transmitting messages across a telecommunication channel, and explicitly sidestepped questions regarding what meaning, if any, the messages might have [35].

How should we fill in the gap that Shannon explicitly introduced? One kind of approach—common in economics, game theory and statistics—begins by assuming an idealized system that pursues some externally assigned goal, usually formulated as the optimization of an objective function, such as utility [37–41], distortion [36] or prediction error [19,42–44]. Semantic information is then defined as information which helps the system to achieve its goal (e.g. information about tomorrow's stock market prices would help a trader increase their economic utility). Such approaches can be quite useful and have lent themselves to important formal developments. However, they have the major shortcoming that they specify the goal of the system *exogenously*, meaning that they are not appropriate for grounding meaning in the *intrinsic* properties of a particular physical system. The semantic information they quantify has meaning for the external scientist who imputes goals to the system, rather than for the system itself.

In biology, the goal of an organism is often considered to be evolutionary success (i.e. the maximization of fitness), which has led to the so-called *teleosemantic* approach to semantic information. Loosely speaking, teleosemantics proposes that a biological trait carries semantic information if the presence of the trait was 'selected for' because, in the evolutionary past, the trait correlated with particular states of the environment [1–7]. To use a well-known example, when a frog sees a small black spot in its visual field, it snaps out its tongue and attempts to catch a fly. This stimulus–response behaviour was selected for, since small black spots in the visual field correlated with the presence of flies and eating flies was good for frog fitness. Thus, a small black spot in the visual field of a frog has semantic information, and refers to the presence of flies.

While in-depth discussion of teleosemantics is beyond the scope of this paper, we note that some of its central features make it deficient for our purposes. First, it is only applicable to physical systems that undergo natural selection. Thus, it is not clear how to apply it to entities like non-living systems, protocells or synthetically designed organisms. Moreover, teleosemantics is 'etiological' [45,46], meaning that it defines semantic information in terms of the past history of a system. Our goal is to develop a theory of semantic information that is based purely on the intrinsic dynamics of a system in a given environment, irrespective of the system's origin and past history.

Finally, another approach to semantic information comes from literature on so-called *autonomous agents* [11,12,14,45–49]. An autonomous agent is a far-from-equilibrium system which actively maintains its own existence within some environment [11–14,25,50–54]. A prototypical example of an autonomous agent is an organism, but in principle, the notion can also be applied to robots [55,56] and other non-living systems [57,58]. For an autonomous agent, self-maintenance is a fundamentally intrinsic goal, which is neither assigned by an external scientist analysing the system, nor based on past evolutionary history.

In order to maintain themselves, autonomous agents must typically observe (i.e. acquire information about) their environment, and then respond in different and 'appropriate' ways. For instance, a chemotactic bacterium senses the direction of chemical gradients in its particular environment and then moves in the direction of those gradients, thereby locating food and maintaining its own existence. In this sense, autonomous agents can be distinguished from 'passive' self-maintaining structures that emerge whenever appropriate boundary conditions are provided, such as Bénard cells [59] and some other well-known non-equilibrium systems.

Research on autonomous agents suggests that information about the environment that is used by an autonomous agent for self-maintenance is intrinsically meaningful [10–14,25,26,48,49,60]. However, until now, such ideas have remained largely informal. In particular, there has been no formal proposal in the autonomous agents literature for quantifying the amount of semantic information possessed by any given physical system, nor for identifying the meaning (i.e. the semantic content) of particular system states.

1.2. Our contribution

We propose a formal, intrinsic definition of semantic information, applicable to any physical system coupled to an external environment, whether a rock, a hurricane, a bacterium, or a sample from an alien planet.[2]

We assume the following set-up: there is a physical world which can be decomposed into two subsystems, which we refer to as 'the system $\mathcal{X}$' and 'the environment $\mathcal{Y}$', respectively. We suppose that at some initial time $t = 0$, the system and environment are jointly distributed according to some initial distribution $p(x_0, y_0)$. They then undergo coupled (possibly stochastic) dynamics until time τ, where τ is some timescale of interest.

Our goal is to define the semantic information that the system has about the environment. To do so, we make use of a **viability function**, a real-valued function which quantifies the system's 'degree of existence' at a given time. While there are several possible ways to define a viability function, in this paper we take inspiration from statistical physics [61–63] and define the viability function as the negative Shannon entropy of the distribution over the states of system $\mathcal{X}$. This choice is motivated by the fact that Shannon entropy provides an upper bound on the probability that the system occupies any small set of 'viable' states [64–67]. We are also motivated by the connection between Shannon entropy and thermodynamics [33,34,68–71], which allows us to connect our framework to results in non-equilibrium statistical physics.

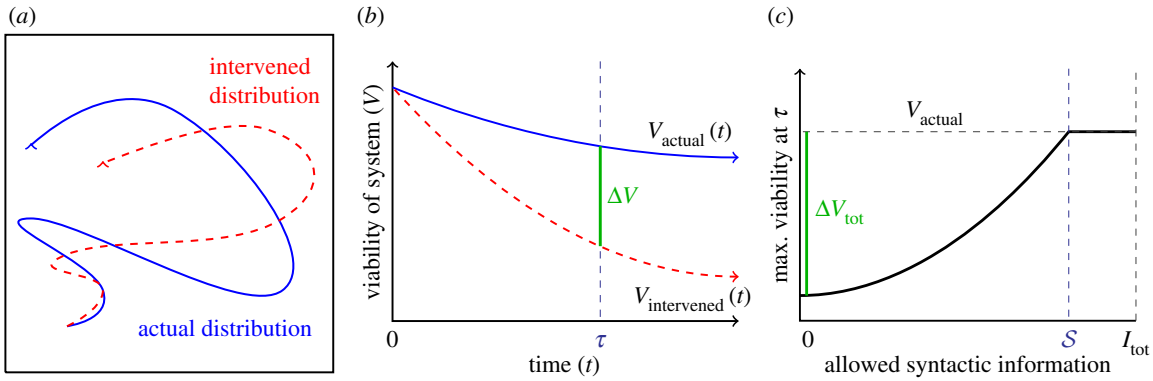

Figure 1. Schematic illustration of our approach to semantic information. (*a*) The trajectory of the actual distribution (within the space of distribution over joint system–environment states) is in blue. The trajectory of the intervened distribution, where some syntactic information between the system and environment is scrambled, is in dashed red. (*b*) The viability function computed for both the actual and intervened trajectories. ΔV indicates the viability difference between actual and intervened trajectories, at some time τ. (*c*) Different ways of scrambling the syntactic information lead to different values of remaining syntactic information and different viability values. The maximum achievable viability at time τ at each level of remaining syntactic information specifies the information/viability curve. The viability value of information, ΔV_{tot}, is the total viability cost of scrambling all syntactic information. The amount of semantic information, $\mathcal{S}$, is the minimum level of syntactic information at which no viability is lost. I_{tot} is the total amount of syntactic information between system and environment. (Online version in colour.)

Further discussion of this viability function, as well as other possible viability functions, is found in §4.

Information theory provides many measures of the syntactic information shared between the system and its environment. For any particular measure of syntactic information, we define semantic information to be *that syntactic information between the system and the environment that causally contributes to the continued existence of the system*, i.e. to maintaining the value of the viability function. To quantify the causal contribution, we define counter-factual **intervened distributions** in which some of the syntactic information between the system and its environment is scrambled. This approach is inspired by the framework of causal interventions [72,73], in which causal effects are measured by counter-factually intervening on one part of a system and then measuring the resulting changes in other parts of the system.

The trajectories of the actual and intervened distributions are schematically illustrated in figure 1*a*. We define the **(viability) value of information** as the difference between the system's viability after time τ under the actual distribution, versus the system's viability after time τ under the intervened distribution (figure 1*b*). A positive difference means that at least some of the syntactic information between the system and environment plays a causal role in maintaining the system's existence. The difference can also be *negative*, which means that the syntactic information decreases the system's ability to exist. This occurs if the system behaves 'pathologically', i.e. it takes the wrong actions given available information (e.g. consider a mutant 'anti-chemotactic' bacterium, which senses the direction of food and then swims away from it).

To make things more concrete, we illustrate our approach using a few examples:

(1) Consider a distribution over rocks (the system) and fields (the environment) over a timescale of $\tau = 1$ year. Rocks tend to stay in a low entropy state for long periods of time due to their very slow dynamics. If we 'scramble the information' between rocks and their environments by swapping rocks between different fields, this will not significantly change the propensity of rocks to disintegrate into (high entropy) dust after 1 year. Since the viability does not change significantly due to the intervention, the viability value of information is very low for a rock.

(2) Consider a distribution over hurricanes (the system) and the summertime Caribbean ocean and atmosphere (the environment), over a timescale of $\tau = 1$ h. Unlike a rock, a hurricane is a genuinely non-equilibrium system which is driven by free energy fluxing from the warm ocean to the cold atmosphere. Nonetheless, if we 'scramble the information' by placing hurricanes in new surroundings that still correspond to warm oceans and cool atmospheres, after 1 h the intervened hurricanes' viability will be similar to that of the non-intervened hurricanes. Thus, like rocks, hurricanes have a low viability value of information.

(3) Consider a distribution over food-caching birds (the system) in the forest (the environment), over a timescale of $\tau = 1$ year. Assume that at $t = 0$ the birds have cached their food and stored the location of the caches in some type of neural memory. If we 'scramble the information' by placing birds in random environments, they will not be able to locate their food and be more likely to die, thus decreasing their viability. Thus, a food-caching bird exhibits a high value of information.

So far, we have spoken of interventions in a rather informal manner. In order to make things rigorous, we require a formal definition of how to transform an actual distribution into an intervened distribution. While we do not claim that there is a single best choice for defining interventions, we propose to use information-theoretic 'coarse-graining' methods to scramble the channel between the system and environment [74–79]. Importantly, such methods allow us to choose different coarse-grainings, which lets us vary the syntactic information that is preserved under different interventions, and the resulting viability of the system at time τ. By considering different interventions, we define a trade-off between the amount of preserved syntactic information versus the resulting viability of the system at time τ. This trade-off is formally represented by an **information/viability curve** (figure 1*c*), which is loosely analogous to the rate-distortion curves in information theory [36].

Note that some intervened distributions may achieve the same viability as the actual distribution but have less

syntactic information. We call the **(viability-) optimal intervention** that intervened distribution which achieves the same viability as the actual distribution while preserving the smallest amount of syntactic information. Using the optimal intervention, we define a number of interesting measures. First, by definition, any further scrambling of the optimal intervention leads to a change in viability of the system, relative to its actual (non-intervened) viability. We interpret this to mean that *all syntactic information in the optimal intervention is semantic information*. Thus, we define the **amount of semantic information** possessed by the system as the amount of syntactic information preserved by the optimal intervention. We show that the amount of semantic information is upper bounded by the amount of syntactic information under the actual distribution, meaning that having non-zero syntactic information is a necessary, but not sufficient, condition for having non-zero semantic information. Moreover, we can decompose the total amount of syntactic information into 'meaningful bits' (the semantic information) and the 'meaningless bits' (the rest), and define the **semantic efficiency** of the system as the ratio of the semantic information to the syntactic information. Semantic efficiency falls between 0 and 1, and quantifies how much the system is 'tuned' to only possess syntactic information which is relevant for maintaining its existence (see also [80]).

Because all syntactic information in the optimal intervention is semantic information, we use the optimal intervention to define the 'content' of the semantic information. The **semantic content** of a particular system state x is defined as the conditional distribution (under the optimal intervention) of the environment's states, given that the system is in state x. The semantic content of x reflects the correlations which are relevant to maintaining the existence of the system, once all other 'meaningless' correlations are scrambled away. To use a previous example, the semantic content for a food-caching bird would include the conditional probabilities of different food-caching locations in the forest, given bird neural states. By applying appropriate 'pointwise' measures of syntactic information to the optimal intervention, we also derive measures of **pointwise semantic information** in particular system states (see §5 for details).

As mentioned, our framework is not tied to one particular measure of syntactic information, but rather can be used to derive different kinds of semantic information from different measures of syntactic information. In §5.1, we consider semantic information derived from the mutual information between the system and environment in the initial distribution $p(x_0, y_0)$, which defines what we call **stored semantic information**. Note that stored semantic information does not measure semantic information which is acquired by ongoing dynamic interactions between system and environment, which is the primary kind of semantic information discussed in the literature on autonomous agents [14]. In §5.2, we derive this kind of dynamically acquired semantic information, which we call **observed semantic information**, from a syntactic information measure called *transfer entropy* [81]. Observed semantic information provides one quantitative definition of **observation**, as dynamically acquired information that is used by a system to maintain its own existence, and allows us to distinguish observation from the mere build-up of syntactic information between physical systems (as generally happens whenever physical systems come into contact). In §5.3, we briefly discuss other possible choices of syntactic information measures, which lead to other measures of semantic information.

Given recent work on the statistical physics of information processing, several of our measures—including value of information and semantic efficiency—can be given thermodynamic interpretations. We review these connections between semantic information and statistical physics in §2, as well as in more depth in §5 when defining stored and observed semantic information.

To summarize, we propose a formal definition of semantic information that is applicable to any physical system. Our definition depends on the specification of a viability function, a syntactic information measure, and a way of producing interventions. We suggest some natural ways of defining these factors, though we have been careful to formulate our approach in a flexible manner, allowing them to be chosen according to the needs of the researcher. Once these factors are determined, our measures of semantic information are defined relative to choice of

(1) the particular division of the physical world into 'the system' and 'the environment';
(2) the timescale τ; and
(3) the initial probability distribution over the system and environment.

These choices specify the particular spatio-temporal scale and state-space regions that interest the researcher, and should generally be chosen in a way to be relevant to the dynamics of the system under study. For instance, if studying semantic information in human beings, one should choose timescales over which information has some effect on the probability of survival (somewhere between $\approx$100 ms, corresponding to the fastest reaction times, and $\approx$100 years). In §6, we discuss how the system/environment decomposition, timescale and initial distribution might be chosen 'objectively', in particular, so as to maximize measures of semantic information. We also discuss how this might be used to automatically identify the presence of agents in physical systems, and more generally the implications of our framework for an intrinsic definition of **autonomous agency** in physical systems.

The rest of the paper is laid out as follows. The next section provides a review of some relevant aspects of non-equilibrium statistical physics. In §3, we provide preliminaries concerning our notation and physical assumptions, while §4 provides a discussion of the viability function. In §5, we state our formal definitions of semantic information and related concepts. Section 6 discusses ways of automatically selecting systems, timescales, and initial distributions so as to maximize semantic information, and implications for a definition of agency. We conclude in §7.

2. Non-equilibrium statistical physics

The connection between the maintenance of low entropy and autonomous agents was first noted when considering the thermodynamics of living systems. In particular, the fact that organisms must maintain themselves in a low entropy state was famously proposed, in an informal manner, by Schrödinger [82], as well as Brillouin [83] and others [84,85]. This had led to an important line of work on quantifying the entropy of various kinds of living matter [86–89].

However, this research did not consider the role of organism–environment information exchanges in maintaining the organism's low entropy state.

Others have observed that organisms not only maintain a low entropy state but also constantly acquire and use information about their environment to do so [52,90–95]. Moreover, it has been suggested that natural selection can drive improvements in the mechanisms that gather and store information about the environment [96]. However, these proposals did not specify how to formally quantify the amount and content of information which contributes to the self-maintenance of any given organism.

Recently, there has been dramatic progress in our understanding of the physics of non-equilibrium processes which acquire, transform, and use information, as part of the development of the so-called thermodynamics of information [34]. It is now well understood that, as a consequence of the Second Law of Thermodynamics, any process that reduces the entropy of a system must incur some thermodynamic costs. In particular, the so-called *generalized Landauer's principle* [69,97,98] states that, given a system coupled to a heat bath at temperature T, any process that reduces the entropy of the system by n bits must release at least $n \cdot k_B T \ln 2$ of energy as heat (alternatively, at most $n \cdot k_B T \ln 2$ of heat can be absorbed by any process that increases entropy by n bits). It has also been shown that in certain scenarios, heat must be generated in order to acquire syntactic information, whether mutual information [34,99–101], transfer entropy [102–106], or other measures [107–111].

Owing to these developments, non-equilibrium statistical physics now has a fully rigorous understanding of 'information-powered non-equilibrium states' [63,99–101,103,112–122], i.e. systems in which non-equilibrium is maintained by the ongoing exchange of information between subsystems. The prototypical case of such situations are 'feedback-control' processes, in which one subsystem acquires information about another subsystem, and then uses this information to apply appropriate control protocols so as to keep itself or the other system out of equilibrium (e.g. Maxwell's demon [121–123], feedback cooling [120], etc.). Information-powered non-equilibrium states differ from the kinds of non-equilibrium systems traditionally considered in statistical physics, which are driven by work reservoirs with (feedback-less) control protocols, or by coupling to multiple thermodynamic reservoirs (e.g. Bénard cells).

Recall that we define our viability functions as the negative entropy of the system. As stated, results from non-equilibrium statistical physics show that both decreasing entropy (i.e. increasing viability) and acquiring syntactic information carries thermodynamic costs, and these costs can be related to each other. In particular, the syntactic information that a system has about its environment will often require some work to acquire. However, the same information may carry an arbitrarily large benefit [124], for instance by indicating the location of a large source of free energy, or a danger to avoid. To compare the benefit and the cost of the syntactic information to the system, below we define the **thermodynamic multiplier** as the ratio between the viability value of the information and the amount of syntactic information. Having a large thermodynamic multiplier indicates that the information that the system has about the environment leads to a large 'bang-per-bit' in terms of viability. As we will see, the thermodynamic multiplier is related to the semantic efficiency of a system: systems with positive value of information and high semantic efficiency tend to have larger thermodynamic multipliers.

3. Preliminaries and physical set-up

We indicate random variables by capital letters, such as X, and particular outcomes of random variables by corresponding lower-case letters, such as x. Lower-case letters $p, q, \ldots$ are also used to refer to probability distributions. Where not clear from context, we use notation like p_X to indicate that p is a distribution of the random variable X. We also use notation like $p_{X,Y}$ for the joint distribution of X and Y, and $p_{X|Y}$ for the conditional distribution of X given Y. We use notation like $p_X p_Y$ to indicate product distributions, i.e. $[p_X p_Y](x, y) = p_X(x) p_Y(y)$ for all x, y.

We assume that the reader is familiar with the basics of information theory [36]. We write $S(p_X)$ for the Shannon entropy of distribution p_X, $I_p(X; Y)$ for the mutual information between random variables X and Y with joint distribution $p_{X,Y}$, and $I_p(X; Y|Z)$ for the conditional mutual information given joint distribution $p_{X,Y,Z}$. We measure information in bits, except where noted.

In addition to the standard measures from information theory, we also use a measure called *transfer entropy* [81]. Given a distribution p over a sequence of paired random variables $(X_0, Y_0), (X_1, Y_1), \ldots, (X_\tau, Y_\tau)$ indexed by timestep $t \in \{0, \ldots, \tau\}$, the transfer entropy from Y to X at timestep t is defined as the conditional mutual information,

$$\mathcal{T}_p(Y_t \to X_{t+1}) = I_p(Y_t; X_{t+1}|X_t). \tag{3.1}$$

Transfer entropy reflects how much knowledge of the state of Y at timestep t reduces uncertainty about the next state of X at the next timestep $t + 1$, conditioned on knowing the state of X at timestep t. It thus reflects 'new information' about Y that is acquired by X at time t.

In our analysis below, we assume that there are two coupled systems, called 'the system $\mathcal{X}$' and 'the environment $\mathcal{Y}$', with state-spaces indicated by X and Y, respectively. The system/environment $X \times Y$ may be isolated from the rest of the universe, or may be coupled to one or more thermodynamic reservoirs and/or work reservoirs. For simplicity, we assume that the joint state space $X \times Y$ is discrete and finite (in physics, such a discrete state space is often derived by coarse-graining an underlying Hamiltonian system [125,126]), though in principle our approach can also be extended to continuous state-spaces. In some cases, $X \times Y$ may also represent a space of coarse-grained macrostates rather than microstates (e.g. a vector of chemical concentrations at different spatial locations), usually under the assumption that local equilibrium holds within each macrostate (see appendix B for an example).

The joint system evolves dynamically from initial time $t = 0$ to final time $t = \tau$. We assume that the decomposition into system/environment remains constant over this time (in future work, it may be interesting to consider time-inhomogeneous decompositions, e.g. for analysing growing systems). In our analysis of observed semantic information in §5.2, we assume for simplicity that the coupled dynamics of $\mathcal{X}$ and $\mathcal{Y}$ are stochastic, discrete-time and first-order Markovian. However, we do not assume that dynamics are time-homogeneous (meaning that, in principle, our framework allows for external driving by the work reservoir).

Other kinds of dynamics (e.g. Hamiltonian dynamics, which are continuous-time and deterministic) can also be considered, though care is needed when defining measures like transfer entropy for continuous-time systems [106].

We use random variables X_t and Y_t to represent the state of $\mathcal{X}$ and $\mathcal{Y}$ at some particular time $t \geq 0$, and random variables $X_{0..\tau} = \langle X_0, \ldots, X_\tau \rangle$ and $Y_{0..\tau} = \langle Y_0, \ldots, Y_\tau \rangle$ to indicate entire trajectories of $\mathcal{X}$ and $\mathcal{Y}$ from time $t = 0$ to $t = \tau$.

4. The viability function

We quantify the 'level of existence' of a given system at any given time with a **viability function** V. Though several viability functions can be considered, in this paper we define the viability function as the negative of the Shannon entropy of the marginal distribution of system $\mathcal{X}$ at time τ,

$$V(p_{X_\tau}) := -S(p_{X_\tau}) = \sum_{x_\tau} p(x_\tau) \log p(x_\tau). \tag{4.1}$$

If the state space of $\mathcal{X}$ represents a set of coarse-grained macrostates, equation (4.1) should be amended to include the contribution from 'internal entropies' of each macrostate (see appendix B for an example).

There are several reasons for selecting negative entropy as the viability function. First, as discussed in §2, results in non-equilibrium statistical physics relate changes of the Shannon entropy of a physical system to thermodynamic quantities like heat and work [33,34,68–71]. These relations allow us to analyse our measures in terms of thermodynamic costs.

The second reason we define viability as negative entropy is that entropy provides an upper bound on the amount of probability that can be concentrated in any small subset of the state space X (for this reason, entropy has been used as a measure of the performance of a controller [61–63]). For us, this is relevant because there is often a naturally defined 'viability set' [64–67,127,128], which is the set of states in which the system $\mathcal{X}$ can continue to perform self-maintenance functions. Typically, the viability set will be a very small subset of the overall state space X. For instance, the total number of ways in which the atoms in an *E. coli* bacterium can be arranged, relative to the number of ways they can be arranged to constitute a living *E. coli*, has been estimated to be of the order of $2^{46\,000\,000}$ [86]. If the entropy of system $\mathcal{X}$ is large and the viability set is small, then the probability that the system state is within the viability set must be small, no matter where that viability set is in X. Thus, maintaining low entropy is a necessary condition for remaining within the viability set. (Appendix A elaborates these points, deriving a bound between Shannon entropy and the probability of the system being within any small subset of its state space.)

At the same time, negative entropy may have some disadvantages as a viability function. Most obviously, a distribution can have low entropy but still assign a low probability to being in a particular viability set. In addition, a system that maintains low entropy over time does not necessarily 'maintain its identity' (e.g. both a rhinoceros and a human have low entropy). Whether this is an advantage or a drawback of the measure depends partly on how the notion of 'self-maintenance' is conceptualized.

There are other ways to define the viability function, some of which address these potential disadvantages of using negative entropy. Given a particular viability set $\mathcal{A} \subseteq X$, a natural definition of the viability function is the probability that the system's state is in the viability set, $p(X_\tau \in \mathcal{A})$. However, this definition requires the viability set to be specified, and in many scenarios we might know that there is a viability set but not be able to specify it precisely. To use a previous example, identifying the viability set of an *E. coli* is an incredibly challenging problem [86].

Alternatively, it is often stated that self-maintaining systems must remain out of thermodynamic equilibrium [11,14,52]. This suggests defining the viability function in a way that captures the 'distance from equilibrium' of system $\mathcal{X}$. One such measure is the Kullback–Leibler divergence (in nats) between the actual distribution over X_τ and the equilibrium distribution of $\mathcal{X}$ at time τ, indicated here by π_{X_τ},

$$D_{\mathrm{KL}}(p_{X_\tau} \| \pi_{X_\tau}). \tag{4.2}$$

This viability function, which is sometimes called 'exergy' or 'availability' in the literature [129,130], has a natural physical interpretation [68]: if the system were separated from environment $\mathcal{Y}$ and coupled to a single heat bath at temperature T, then up to $k_{\mathrm{B}}T \cdot D_{\mathrm{KL}}(p_{X_\tau} \| \pi_{X_\tau})$ work could be extracted by bringing the system from p_{X_τ} to π_{X_τ}.

Unfortunately, there are difficulties in using equation (4.2) as the viability function in the general case. In statistical physics, the equilibrium distribution is defined as a stationary distribution in which all probability fluxes vanish. Since the system $\mathcal{X}$ is open (it is coupled to the environment $\mathcal{Y}$, and possibly multiple thermodynamic reservoirs), such an equilibrium distribution will not exist in the general case, and equation (4.2) may be undefined. For instance, a Bénard cell, a well-known non-equilibrium system which is coupled to both hot and cold thermal reservoirs [59], will evolve to a *non-equilibrium* stationary distribution, in which probability fluxes do not vanish. While it is certainly true that a Bénard cell is out of thermodynamic equilibrium, one cannot quantify 'how far' from equilibrium it is by using equation (4.2).

In principle, it is possible to quantify the 'amount of non-equilibrium' without making reference to an equilibrium distribution, in particular, by measuring the amount of probability flux in a system (e.g. instantaneous entropy production [131,132] or the norm of the probability fluxes [133,134]). However, there is not necessarily a clear relationship between the amount of probability flux and the capacity of a system to carry out self-maintenance functions [135]. We leave exploration of these alternative viability functions for future work.

It is important to re-emphasize that, in our framework, the viability function is exogenously determined by the scientist analysing the system, rather than being a purely endogenous characteristic of the system. At first glance, our approach may appear to suffer some of the same problems as do approaches that define semantic information in terms of an exogenously specified utility function (see the discussion in §1.1). However, there are important differences between a utility function and a viability function. First, we require that a viability function is well defined for *any* physical system, whether a rock, a human, a city, a galaxy; utility functions, on the other hand, are generally scenario-specific and far from universal. Furthermore, given an agent with an exogenously defined utility function operating in a time-extended scenario, maintaining existence is almost always a necessary (though usually implicit) condition for high utility. A

reasonably chosen viability function should capture this minimal, universal component of nearly all utility functions. Finally, unlike utility functions, in principle, it may be possible to derive the viability function in some objective way (e.g. in terms of the attractor landscape of the coupled system–environment dynamics [64,128]).

5. Semantic information via interventions

As described above, we quantify semantic information in terms of the amount of syntactic information which contributes to the ability of the system to continue existing.

We use the term **actual distribution** to refer to the original, unintervened distribution of trajectories of the joint system–environment over time $t = 0$ to $t = \tau$, which will usually be indicated with the symbol p. Our goal is to quantify how much semantic information the system has about the environment under the actual distribution. To do this, we define a set of counter-factual **intervened distributions** over trajectories, which are similar to the actual distribution except that some of syntactic information between system and environment is scrambled, and which will usually be indicated with some variant of the symbol $\hat{p}$. We define measures of semantic information by analysing how the viability of the system at time τ changes between the actual and the intervened distributions.

Information theory provides many different measures of syntactic information between the system and environment, each of which requires a special type of intervention, and each of which gives rise to a particular set of semantic information measures. In this paper, we focus on two types of syntactic information. In §5.1, we consider **stored semantic information**, which is defined by scrambling the mutual information between system and environment in the actual initial distribution p_{X_0,Y_0}, while leaving the dynamics unchanged. In §5.2, we instead consider **observed semantic information**, which is defined via a 'dynamic' intervention in which we keep the initial distribution the same but change the dynamics so as to scramble the transfer entropy from the environment to the system. Observed semantic information identifies semantic information that is acquired by dynamic interactions between the system and environment, rather than present in the initial mutual information. An example of observed semantic information is exhibited by a chemotactic bacterium, which makes ongoing measurements of the direction of food in its environment, and then uses this information to move towards food. In §5.3, we briefly discuss other possible measures of semantic information.

5.1. Stored semantic information

5.1.1. Overview

Stored semantic information is derived from the mutual information between system and environment at time $t = 0$. This mutual information can be written as

$$I_p(X_0,Y_0) = \sum_{x_0,y_0} p(x_0,y_0) \log \frac{p(x_0,y_0)}{p(x_0)p(y_0)}. \tag{5.1}$$

Mutual information achieves its minimum value of 0 if and only if X_0 and Y_0 are statistically independent under p, i.e. when $p_{X_0,Y_0} = p_{X_0} p_{Y_0}$. Thus, we first consider an intervention that destroys all mutual information by transforming the actual initial distribution p_{X_0,Y_0} to the product initial distribution,

$$p_{X_0,Y_0} \mapsto \hat{p}^{\text{full}}_{X_0,Y_0} := p_{X_0} p_{Y_0}. \tag{5.2}$$

(We use the superscript 'full' to indicate that this is a 'full scrambling' of the mutual information.)

To compute the **viability value** of stored semantic information at $t = 0$, we run the coupled system–environment dynamics starting from both the actual initial distribution p_{X_0,Y_0} and the intervened initial distribution $\hat{p}^{\text{full}}_{X_0,Y_0}$, and then measure the difference in the viability of the system at time τ,

$$\Delta V^{\text{stored}}_{\text{tot}} := V(p_{X_\tau}) - V(\hat{p}^{\text{full}}_{X_\tau}). \tag{5.3}$$

For the particular viability function we are considering (negative entropy), the viability value is

$$\Delta V^{\text{stored}}_{\text{tot}} = S(\hat{p}^{\text{full}}_{X_\tau}) - S(p_{X_\tau}). \tag{5.4}$$

Equation (5.3) measures the difference of viability under the 'full scrambling', but does not specify which part of the mutual information actually causes this difference. To illustrate this issue, consider a system in an environment where food can be in one of two locations with 50% probability each, and the system starts at $t = 0$ with perfect information about the food location. Imagine that system's viability depends upon it finding and eating the food. Now suppose that the system also has 1000 bits of mutual information about the state of the environment which does not contribute in any way to the system's viability. In this case, the initial mutual information will be 1001 bits, though only 1 bit (the location of the food) is 'meaningful' to the system, in that it affects the system's ability to maintain high viability.

In order to find that part of the mutual information which is meaningful, we define an entire set of 'partial' interventions (rather than just considering the single 'full' intervention mentioned above). We then find the partial intervention which destroys the most syntactic information while leaving the viability unchanged, which we call the **(viability-) optimal intervention**. The optimal intervention specifies which part of the mutual information is meaningless, in that it can be scrambled without affecting viability, and which part is meaningful, in the sense that it must be preserved in order to achieve the actual viability value. For the example mentioned in the previous paragraph, the viability-optimal intervention would preserve the 1 bit of information concerning the location of the food, while scrambling away the remaining 1000 bits.

Each partial interventions in the set of possible partial interventions is induced by a particular 'coarse-graining function'. First, consider the actual conditional probability of system given environment at $t = 0$, $p_{X_0|Y_0}$, as a communication channel over which the system acquires information from its environment. To define each partial intervention, we coarse-grain this communication channel $p_{X_0|Y_0}$ using a coarse-graining function $\phi(y)$, which specifies which distinctions the system can make about the environment. Formally, the intervened channel from Y_0 to X_0 induced by ϕ, indicated as $\hat{p}^{\phi}_{X_0|Y_0}$, is taken to be the

actual conditional probability of system states X_0 given coarse-grained environments $\phi(Y_0)$,

$$\hat{p}^{\phi}(x_0|y_0) := p(x_0|\phi(y_0)) = \frac{\sum_{y_0':\phi(y_0')=\phi(y_0)} p(x_0,y_0')}{\sum_{y_0':\phi(y_0')=\phi(y_0)} p(y_0')}. \quad (5.5)$$

We then define the intervened joint distribution at $t = 0$ as $\hat{p}^{\phi}_{X_0,Y_0} := \hat{p}^{\phi}_{X_0|Y_0} p_{Y_0}$. Under the intervened distribution $\hat{p}^{\phi}_{X_0,Y_0}$, X_0 is conditionally independent of Y_0 given $\phi(Y_0)$, and any two states of the environment y_0 and y_0' which have $\phi(y_0) = \phi(y_0')$ will be indistinguishable from the point of view of the system. Said differently, X_0 will only have information about $\phi(Y_0)$, not Y_0 itself, and it can be verified that $I_{\hat{p}^{\phi}}(X_0; Y_0) = I_p(X_0; \phi(Y_0))$. In the information-theory literature, the coarse-grained channel $\hat{p}^{\phi}_{X_0|Y_0}$ is sometimes called a 'Markov approximation' of the actual channel $p_{X_0|Y_0}$ [77], which is itself a special case of the so-called channel pre-garbling or channel input-degradation [77–79]. Pre-garbling is a principled way to destroy part of the information flowing across a channel, and has important operationalizations in terms of coding and game theory [78].

So far we have left unspecified how the coarse-graining function ϕ is chosen. In fact, one can choose different ϕ, in this way inducing different partial interventions. The 'most conservative' intervention corresponds to any ϕ which is a one-to-one function of Y, such as the identity map $\phi(y) = y$. In this case, one can use equation (5.5) to verify that the intervened channel from Y_0 to X_0 will be the same as the actual channel, and the intervention will have no effect. The 'least conservative' intervention occurs when ϕ is a constant function, such as $\phi(y) = 0$. In this case, the intervened distribution will be the 'full scrambling' of equation (5.2), for which $I_{\hat{p}^{\phi}}(X_0; Y_0) = 0$. We use Φ to indicate the set of all possible coarse-graining functions (without loss of generality, we can assume that each element of this set is $\phi: Y \to Y$).

We are now ready to define our remaining measures of stored semantic information. We first define the **information/viability curve** as the maximal achievable viability at time τ under any possible intervention,

$$\mathcal{D}_{\text{stored}}(R) := \max_{\phi\in\Phi} V(\hat{p}^{\phi}_{X_\tau}) \quad \text{s.t.} \quad I_{\hat{p}^{\phi}}(X_0, Y_0) = R,$$

where R indicates the amount of mutual information that is preserved. (Note that $\mathcal{D}_{\text{stored}}(R)$ is undefined for values of R when there is no function ϕ such that $I_{\hat{p}^{\phi}}(X_0,Y_0) = R$.) $\mathcal{D}_{\text{stored}}(R)$ is the curve schematically diagrammed in figure 1*c*.

We define the **(viability-) optimal intervention** $\hat{p}^{\text{opt}}_{X_0,Y_0}$ as the intervention that achieves the same viability value as the actual distribution while having the smallest amount of syntactic information,

$$\hat{p}^{\text{opt}}_{X_0,Y_0} \in \underset{\hat{p}^{\phi}:\phi\in\Phi}{\text{argmin}}\, I_{\hat{p}^{\phi}}(X_0, Y_0) \quad \text{s.t.} \quad V(\hat{p}^{\phi}_{X_\tau}) = V(p_{X_\tau}). \quad (5.6)$$

By definition, any further scrambling of $\hat{p}^{\text{opt}}_{X_0,Y_0}$ would change system viability, meaning that in $\hat{p}^{\text{opt}}_{X_0,Y_0}$ all remaining mutual information is meaningful. Therefore, we define the **amount of stored semantic information** as the mutual information in the optimal intervention,

$$\mathcal{S}_{\text{stored}} := I_{\hat{p}^{\text{opt}}}(X_0, Y_0). \quad (5.7)$$

While the value of information $\Delta V^{\text{stored}}_{\text{tot}}$ can be positive or negative, the amount of stored semantic information is always non-negative. Moreover, stored semantic information reflects the number of bits that play a causal role in determining the viability of the system at time τ, regardless in whether they cause it to change positively or negatively.

Since the actual distribution p_{X_0,Y_0} is part of the domain of the minimization in equation (5.6) (it corresponds to any ϕ which is one-to-one), the amount of stored semantic information $I_{\hat{p}^{\text{opt}}}(X_0,Y_0)$ must be less than the actual mutual information $I_p(X_0, Y_0)$. We define the **semantic efficiency** as the ratio of the stored semantic information to the overall syntactic information,

$$\eta_{\text{stored}} := \frac{\mathcal{S}_{\text{stored}}}{I_p(X_0, Y_0)} \in [0, 1]. \quad (5.8)$$

Semantic efficiency measures what portion of the initial mutual information between the system and environment causally contributes to the viability of the system at time τ.

5.1.2. Pointwise measures

As mentioned, the optimal intervention only contains semantic information, i.e. only information which affects the viability of the system at time τ. We use this to define the **pointwise semantic information** of individual states of the system and environment in terms of 'pointwise' measures of mutual information [136] under $\hat{p}^{\text{opt}}$,

$$\mathcal{S}_{\text{stored}}(x_0; y_0) := \log \frac{\hat{p}^{\text{opt}}(x_0, y_0)}{\hat{p}^{\text{opt}}(x_0)\,\hat{p}^{\text{opt}}(y_0)}. \quad (5.9)$$

We similarly define the **specific semantic information** in system state x_0 as the 'specific information' [137] about Y given x_0,

$$\mathcal{S}_{\text{stored}}(x_0; Y_0) = \sum_{y_0} \hat{p}^{\text{opt}}(y_0|x_0) \log \frac{\hat{p}^{\text{opt}}(y_0|x_0)}{\hat{p}^{\text{opt}}(y_0)}. \quad (5.10)$$

These measures quantify the extent to which a system state x_0, and a system–environment state x_0, y_0, carry correlations which causally affect the system's viability at $t = \tau$. Note that the specific semantic information, equation (5.10), and overall stored semantic information, equation (5.7), are expectations of the pointwise semantic information, equation (5.9).

Finally, we define the **semantic content** of system state x_0 as the conditional distribution $\hat{p}^{\text{opt}}(y_0|x_0)$ over all $y_0 \in Y$. The semantic content of x_0 reflects the precise set of correlations between x_0 and the environment at $t = 0$ that causally affect the system's viability at time τ.

It is important to note that the optimal intervention may not be unique, i.e. there might be multiple minimizers of equation (5.6). In case there are multiple optimal interventions, each optimal intervention will have its own measures of semantic content, and its own measures of pointwise and specific semantic information. The non-uniqueness of the optimal intervention, if it occurs, indicates that the system possesses multiple *redundant* sources of semantic information, any one of which is sufficient to achieve the actual viability value at time τ. A prototypical example is when the system has information about multiple sources of food which all provide the same viability benefit, and where the system can access at most one food source during $t \in [0, \tau]$.

5.1.3. Thermodynamics

In this section, we use ideas from statistical physics to define the **thermodynamic multiplier** of stored semantic information.

This measure compares the physical costs to the benefits of system–environment mutual information.

We begin with a simple illustrative example. Imagine a system coupled to a heat bath at temperature T, as well as an environment which contains a source of 10^6 J of free energy (e.g. a hamburger) in one of two locations (A or B), with 50% probability each. Assume that the system only has time to move to only one of these locations during the interval $t \in [0, \tau]$. We now consider two scenarios. In the first, the system initially has 1 bit of information about the location of the hamburger, which will generally cost at least $k_B T \ln 2$ of work to acquire. The system can use this information to move to the hamburger's location and then extract 10^6 J of free energy. In the second scenario, the system never acquires the 1 bit of information about the hamburger location, and instead starts from the 'fully scrambled' distribution $\hat{p}^{\text{full}}_{X_0,Y_0} = p_{X_0} p_{Y_0}$ (equation (5.2)). By not acquiring the 1 bit of information, the system can save $k_B T \ln 2$ of work, which could be used at time τ to decrease its entropy (i.e. increase its viability) by 1 bit. However, because the system has no information about the hamburger location, it only finds the hamburger 50% of the time, thereby missing out on 0.5×10^6 J of free energy on average. This amount of lost free energy could have been used to decrease the system's entropy by $0.5 \times 10^6/(k_B T \ln 2)$ bits at time $t = \tau$. At typical temperatures, $0.5 \times 10^6/(k_B T \ln 2) \gg 1$, meaning that the benefit of having the bit of information about the hamburger location far outweighs the cost of acquiring that bit.

To make this argument formal, imagine a physical 'measurement' process that transforms the fully scrambled system–environment distribution $\hat{p}^{\text{full}}_{X_0,Y_0} = p_{X_0} p_{Y_0}$ to the actual joint distribution p_{X_0,Y_0}. Assume that during the course of this process, the interaction energy between $\mathcal{X}$ and $\mathcal{Y}$ is negligible and that a heat bath at temperature T is available. The minimum amount of work required by any such measurement process [34,100] is $k_B T \ln 2$ times the change of system-environment entropy in bits, $\Delta S = [S(p_{X_0}) + S(p_{Y_0})] - S(p_{X_0,Y_0}) = I_p(X_0; Y_0)$. We take this minimum work,

$$W_{\text{min}} = k_B T \ln 2 \cdot I_p(X_0; Y_0), \tag{5.11}$$

to be the cost of acquiring the mutual information. If this work were not spent acquiring the initial mutual information, it could have been used at time τ to decrease the entropy of the system, and thereby increase its viability, by $I_p(X_0; Y_0)$ (again ignoring energetic considerations).

The benefit of the mutual information is quantified by the viability value $\Delta V^{\text{stored}}_{\text{tot}}$, which reflects the difference in entropy at time $t = \tau$ when the system is started in its actual initial distribution p_{X_0,Y_0} versus the fully scrambled initial distribution $\hat{p}^{\text{full}}_{X_0,Y_0} = p_{X_0} p_{Y_0}$, as in equation (5.4).

Combining, we define the **thermodynamic multiplier** of stored semantic information, κ_{stored}, as the benefit/cost ratio of the mutual information,[3]

$$\kappa_{\text{stored}} = \frac{\Delta V^{\text{stored}}_{\text{tot}}}{I_p(X_0; Y_0)} = \frac{S(\hat{p}^{\text{full}}_{X_\tau}) - S(p_{X_\tau})}{I_p(X_0; Y_0)}. \tag{5.12}$$

The thermodynamic multiplier quantifies the 'bang-per-bit' that the syntactic information provides to the system, and provides a way to compare the ability of different systems to use information to maintain their viability high. $\kappa_{\text{stored}} > 1$ means that the benefit of the information outweighs its cost. The thermodynamic multiplier can also be related to semantic efficiency, equation (5.8), via

$$\kappa_{\text{stored}} = \eta_{\text{stored}} \frac{\Delta V^{\text{stored}}_{\text{tot}}}{\mathcal{S}_{\text{stored}}}.$$

If the value of information is positive, then having a low semantic efficiency η_{stored} translates into having a low thermodynamic multiplier. Thus, there is a connection between 'paying attention to the right information', as measured by semantic efficiency, and being thermodynamically efficient.

It is important to emphasize that we do not claim that the system *actually* spends $k_B T \ln 2 \cdot I_p(X_0; Y_0)$ of work to acquire the mutual information in p_{X_0,Y_0}. The actual cost could be larger, or it could be paid by the environment $\mathcal{Y}$ rather than the system, or by an external agent that prepares the joint initial condition of $\mathcal{X}$ and $\mathcal{Y}$, etc. Instead, the above analysis provides a way to compare the thermodynamic cost of acquiring the initial mutual information to the viability benefit of that mutual information. In situations where the actual cost of measurements performed by a system can be quantified (e.g. by counting the number of used ATPs), one could define the thermodynamic multiplier in terms of this actual cost.

Finally, we also emphasize that we ignore all energetic considerations in the above operationalization of the thermodynamic multiplier, in part by assuming a negligible interaction energy between system and environment. We have similarly ignored all energetic consequences in our analysis of interventions, as described above. It is not clear whether this approach is always justified. For instance, imagine that the system and environment have a large interaction energy at $t = 0$. In this case, a 'measurement process' that performs the transformation $p_{X_0} p_{Y_0} \mapsto p_{X_0,Y_0}$—or alternatively an 'intervention process' that performs the full scrambling $p_{X_0,Y_0} \mapsto p_{X_0} p_{Y_0}$—may involve a very large (positive or negative) change in expected energy. Assuming the system–environment Hamiltonian is specified, one may consider defining a thermodynamic multiplier that takes into account changes in expected energy. Furthermore, one may also consider defining interventions in a way that obeys energetic constraints, so that interventions scramble information without injecting or extracting a large amount of energy into the system and environment. Exploring such extensions remains for future work.

5.1.4. Example: food-seeking agent

We demonstrate our framework using a simple model of a food-seeking agent. In this model, the environment $\mathcal{Y}$ contains food in one of five locations (initially uniformly distributed). The agent $\mathcal{X}$ can also be located in one of these five locations, and has internal information about the location of the food (i.e. its 'target'). The agent always begins in location three (the middle of the world). Under the actual initial distribution, the agent has exact information about the location of the food. In each timestep, the agent moves towards its target and if it ever finds itself within one location of the food, it eats the food. If the agent does not eat food for a certain number of timesteps, it enters a high-entropy 'death' macrostate, which it can only exit with an extremely small probability (of the order of $\approx 10^{-34}$).

Figure 2 shows the results for timescale $\tau = 5$. The initial mutual information is $\log_2 5 \approx 2.32$ bits, corresponding to the five possible locations of the food. However, the total amount of stored semantic information is only ≈ 1.37 bits, giving a

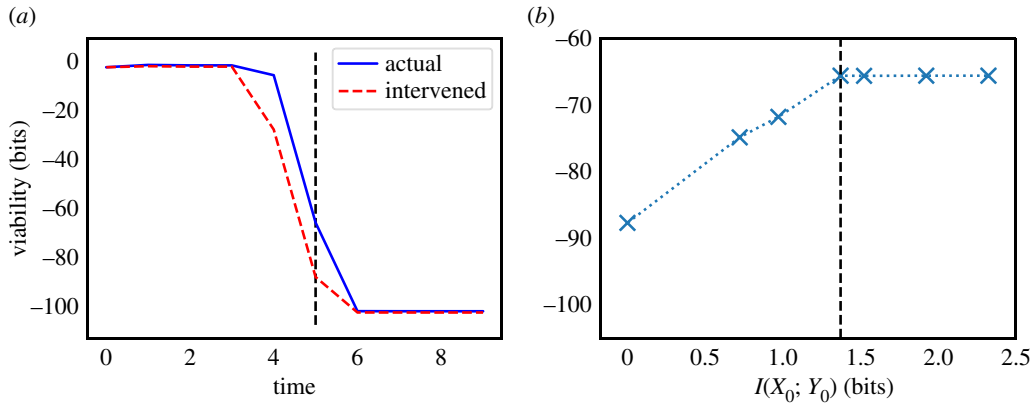

Figure 2. Illustration of our approach using a simple model of a food-seeking agent. (*a*) We plot viability values over time under both the actual and (fully scrambled) intervened distributions. The vertical dashed line corresponds to our timescale of interest ($\tau = 5$ timesteps). (*b*) We plot the information/viability curve for $\tau = 5$ ($\times$'s are actual points on the curve, dashed line is interpolation). The vertical dashed line indicates the amount of stored semantic information. See text for details. (Online version in colour.)

semantic efficiency of $\eta_{\text{stored}} \approx 0.6$. This occurs because if the food is initially in locations $\{2, 3, 4\}$, the agent is close enough to eat it immediately. From the point of view of the agent, differences between these three locations are 'meaningless' and can be scrambled with no loss of viability. Formally, the (unique) optimal intervention $\hat{p}^{\text{opt}}$ is induced by the following coarse-graining function:

$$\phi(y_0) = \begin{cases} 1 & \text{if } y_0 = 1 \\ 3 & \text{if } y_0 \in \{2, 3, 4\} \\ 5 & \text{if } y_0 = 5 \end{cases}$$

which is neither one-to-one nor a constant function (thus, it is a strictly partial intervention). The value of information is $\Delta V_{\text{tot}}^{\text{stored}} \approx 22.1$ bits, giving a thermodynamic multiplier of $\kappa_{\text{stored}} \approx 9.5$ (the food is 'worth' about 9.5 times more than the possible cost of acquiring information about its location).

In appendix B, we describe this model in detail, as well as a variation in which the system moves *away* from food rather than towards it, and thus has negative value of information. A Python implementation can be found at https://github.com/artemyk/semantic_information/.

5.2. Observed semantic information

To identify dynamically acquired semantic information, which we call **observed semantic information**, we define interventions in which we perturb the dynamic flow of syntactic information from environment to system, without modifying the initial system–environment distribution. While there are many ways of quantifying such information flow, here we focus on a widely used measure called transfer entropy [81]. Transfer entropy has several attractive features: it is directed (the transfer entropy from environment to system is not necessarily the same as the transfer entropy from system to environment), it captures common intuitions about information flow, and it has undergone extensive study, including in non-equilibrium statistical physics [102–106].

Observed semantic information can be illustrated with the following example. Imagine a system coupled to an environment in which the food can be in one of two locations (A or B), each of which occurs with 50% probability. At $t = 0$, the system has no information about the location of the food, but the dynamics are such that it acquires and internally stores this location in transitioning from $t = 0$ to $t = 1$. If we intervene and 'fully' scramble the transfer entropy, then in transitioning from $t = 0$ to $t = 1$ the system would find itself 'measuring' location A and B with 50% probability each, independently of the actual food location. Thus, if the system used its measurements to move towards food, it would find itself finding food with only 50% probability, and its viability would suffer. In this case, the transfer entropy from environment to system would contain observed semantic information.

Our approach is formally and conceptually similar to the one used to define stored semantic information (§5.1), and we proceed in a more cursory manner.

The transfer entropy from Y to X over $t \in [1..\tau]$ under the actual distribution can be expressed as a sum of conditional mutual information terms (see equation (3.1)),

$$\sum_{t=0}^{\tau-1} \mathcal{T}_p(Y_t \to X_{t+1}) = \sum_{t=0}^{\tau-1} I_p(X_{t+1}; Y_t | X_t). \tag{5.13}$$

Note that the overall stochastic dynamics of the system and environment at time t can be written as $p_{X_{t+1},Y_{t+1}|X_t,Y_t} = p_{X_{t+1}|X_t,Y_t}\, p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$, where $p_{X_{t+1}|X_t,Y_t}$ represents the response of the system to the previous state of itself and the environment, while $p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$ represents the response of the environment to the previous state of itself and the system, as well as the current state of the system. Observe that the conditional mutual information at time t depends only on $p_{X_{t+1}|X_t,Y_t}$, not on $p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$. Thus, we define a set of partial interventions in which we partially scramble the conditional distribution $p_{X_{t+1}|X_t,Y_t}$, while keeping the conditional distribution $p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$ undistributed. This ensures that our interventions only perturb the information flow from the environment to the system, and not vice versa.[4]

We now define our intervention procedure formally. As mentioned, the conditional distribution $p_{X_{t+1}|X_t,Y_t}$ specifies how information flows from the environment to the system at time t. Each partial intervention is defined by using a coarse-graining function $\phi(y)$, which is used to produce an intervened 'coarse-grained' version of this conditional distribution at all times t. The intervened conditional distribution induced by ϕ at time t, indicated as $\hat{p}^{\phi}_{X_{t+1}|X_t,Y_t}$, is defined to

be the same as the conditional distribution of X_{t+1} given X_t and the coarse-grained environment $\phi(Y_t)$,

$$\hat{p}^{\phi}(x_{t+1}|x_t, y_t) := \hat{p}^{\phi}(x_{t+1}|x_t, \phi(y_t)) \tag{5.14}$$

$$= \frac{\sum_{y_t': \phi(y_t')=\phi(y_t)} p(x_{t+1}|x_t, y_t')\, \hat{p}^{\phi}(x_t, y_t')}{\sum_{y_t': \phi(y_t')=\phi(y_t)} \hat{p}^{\phi}(x_t, y_t')}. \tag{5.15}$$

Note that this definition depends on both the actual dynamics, $p_{X_{t+1}|X_t, Y_t}$ and on the intervened system–environment distribution at time t, $\hat{p}^{\phi}_{X_t, Y_t}$. Under the intervened distribution, X_{t+1} is guaranteed to only have conditional information about $\phi(Y_t)$, not Y_t itself; formally, one can verify that $I_{\hat{p}^{\phi}}(X_{t+1}; Y_t|X_t) = I_p(X_{t+1}; \phi(Y_t)|X_t)$. These definitions are largely analogous to the ones defined for stored semantic information, and the reader should consult that section for more motivation of such coarse-graining procedures.

Under the intervened distribution, the joint system–environment dynamics at time t are computed as $\hat{p}^{\phi}_{X_{t+1}, Y_{t+1}|X_t, Y_t} := \hat{p}^{\phi}_{X_{t+1}|X_t, Y_t} p_{Y_{t+1}|X_t, Y_t, X_{t+1}}$. Then, the overall intervened dynamical trajectory from time $t = 0$ to $t = \tau$, indicated by $\hat{p}^{\phi}_{X_{0..\tau}, Y_{0..\tau}}$, is computed via the following iterative procedure:

(1) At $t = 0$, the intervened system–environment distribution is equal to the actual one, $\hat{p}^{\phi}_{X_0, Y_0} = p_{X_0, Y_0}$.
(2) Using $\hat{p}^{\phi}_{X_t, Y_t}$ and the above definitions, compute $\hat{p}^{\phi}_{X_{t+1}, Y_{t+1}|X_t, Y_t}$.
(3) Using $\hat{p}^{\phi}_{X_{t+1}, Y_{t+1}|X_t, Y_t}$, update $\hat{p}^{\phi}_{X_{0..t}, Y_{0..t}}$ to $\hat{p}^{\phi}_{X_{0..t+1}, Y_{0..t+1}}$.
(4) Set $t \leftarrow t + 1$ and repeat the above steps if $t < \tau$.

We define Φ to be set of all possible coarse-graining functions. By choosing different coarse-graining functions $\phi \in \Phi$, we can produce different partial interventions. One can verify from equation (5.14) that the intervened distribution $\hat{p}^{\phi}_{X_{0..\tau}, Y_{0..\tau}}$ will equal to the actual $p_{X_{0..\tau}, Y_{0..\tau}}$ whenever ϕ is a one-to-one function. When ϕ is a constant function, the intervened distribution will be a 'fully scrambled' one, in which X_{t+1} is conditionally independent of Y_t given X_t for all times t,

$$\hat{p}^{\text{full}}_{X_{t+1}|X_t, Y_t} = \hat{p}^{\phi}_{X_{t+1}|X_t}. \tag{5.16}$$

In this case, the transfer entropy at every time step will vanish.

We are now ready to define our measures of observed semantic information, which are analogous to the definition in §5.1, but now defined for transfer entropy rather than initial mutual information. The **viability value of transfer entropy** is the difference in viability at time τ between the actual distribution and the fully scrambled distribution,

$$\Delta V^{\text{observed}}_{\text{tot}} = V(p_{X_\tau}) - V(\hat{p}^{\text{full}}_{X_\tau}), \tag{5.17}$$

where $\hat{p}^{\text{full}}_{X_\tau}$ is the distribution over X at time τ induced by the fully scrambled intervention. The viability value measures the overall impact of scrambling all transfer entropy on viability. We define **information/viability curve** as the maximal achievable viability for any given level of preserved transfer entropy,

$$\mathcal{D}_{\text{observed}}(R) := \max_{\phi} V(\hat{p}^{\phi}_{X_\tau}) \quad \text{s.t.} \sum_{t=0}^{\tau-1} \mathcal{T}_{\hat{p}^{\phi}}(Y_t \to X_{t+1}) = R.$$

The **(viability-) optimal intervention** $\hat{p}^{\text{opt}}{}_{X_{0..\tau}, Y_{0..\tau}}$ is defined as the intervened distribution that achieves the same viability value as the actual distribution while having the smallest amount of transfer entropy,

$$\hat{p}^{\text{opt}}{}_{X_{0..\tau}, Y_{0..\tau}} \in \operatorname*{argmin}_{\hat{p}^{\phi}: \phi \in \Phi} \sum_{t=0}^{\tau-1} \mathcal{T}_{\hat{p}^{\phi}}(Y_t \to X_{t+1}) \quad \text{s.t.}\ V(\hat{p}^{\phi}_{X_\tau}) = V(p_{X_\tau}). \tag{5.18}$$

Under the optimal intervention, $\hat{p}^{\text{opt}}{}_{X_{0..\tau}, Y_{0..\tau}}$, all meaningless bits of transfer entropy are scrambled while all remaining transfer entropy is meaningful. We use this to define the **amount of observed semantic information** as the amount of transfer entropy under the optimal intervention,

$$\mathcal{S}_{\text{observed}} = \sum_{t=0}^{\tau-1} \mathcal{T}_{\hat{p}^{\text{opt}}}(Y_t \to X_{t+1}). \tag{5.19}$$

Finally, we define the **semantic efficiency** of observed semantic information as the ratio of the amount of observed semantic information to the overall transfer entropy,

$$\eta_{\text{observed}} := \frac{\mathcal{S}_{\text{observed}}}{\sum_{t=0}^{\tau-1} \mathcal{T}_p(Y_t \to X_{t+1})} \in [0, 1].$$

Semantic efficiency quantifies which portion of transfer entropy determines the system's viability at time τ. It is non-negative due the non-negativity of transfer entropy. It is upper bounded by 1 because the actual distribution over system–environment trajectories, $p_{X_{0..\tau}, Y_{0..\tau}}$, is part of the domain of the minimization in equation (5.17) (corresponding to any ϕ which is a one-to-one function), thus the amount of observed semantic information $\mathcal{S}_{\text{observed}}$ will always be less than the actual amount of transfer entropy $\mathcal{T}_p(Y_t \to X_{t+1})$.

We now use the fact that $\hat{p}^{\text{opt}}$ contains only meaningful bits of transfer entropy to define both the semantic content and pointwise measures of observed semantic information. Note that transfer entropy at time t can be written as

$$\mathcal{T}_{\hat{p}^{\text{opt}}}(Y_t \to X_{t+1}) = \sum_{x_t, y_t, x_{t+1}} \hat{p}^{\text{opt}}(x_t, y_t, x_{t+1}) \log \frac{\hat{p}^{\text{opt}}(y_t|x_t, x_{t+1})}{\hat{p}^{\text{opt}}(y_t|x_t)}.$$

We define the **semantic content of the transition** $x_t \mapsto x_{t+1}$ as the conditional distribution $\hat{p}^{\text{opt}}(y_t|x_t, x_{t+1})$ for all $y_t \in Y$. This conditional distribution captures only those correlations between (x_t, x_{t+1}) and Y_t that contribute to the system's viability. Similarly, we define **pointwise observed semantic information** using 'pointwise' measures of transfer entropy [138,139] under $\hat{p}^{\text{opt}}$. In particular, the pointwise observed semantic information for the transition $x_t \mapsto x_{t+1}$ can be defined as

$$\mathcal{S}_{\text{observed}}(y_t|x_t, x_{t+1}) := \log \frac{\hat{p}^{\text{opt}}(y_t|x_t, x_{t+1})}{\hat{p}^{\text{opt}}(y_t|x_t)}.$$

It is of interest to define the **thermodynamic multiplier** for observed semantic information, so as to compare the viability value of transfer entropy to the cost of acquiring that transfer entropy. However, there are different ways of quantifying the thermodynamic cost of acquiring transfer entropy, which depend on the particular way that the measurement process is operationalized [102–106]. Because this thermodynamic analysis is more involved than the one for stored semantic information, we leave it for future work.

5.3. Other kinds of semantic information

We have discussed semantic information defined relative to two measures of syntactic information: mutual information at $t = 0$, and transfer entropy incurred over the course of $t \in [0..\tau]$. In future work, a similar approach can be used to define the semantic information relative to other measures of syntactic information. For example, one could consider the semantic information in the transfer entropy from the system to the environment, which would reflect how much 'observations by the environment' affect the viability of the system (an example of a system with this kind of semantic information is a human coupled to a so-called 'artificial pancreas' [140], a medical device which measures a person's blood glucose and automatically delivers necessary levels of insulin). Alternatively, one might evaluate how mutual information (or transfer entropy, etc.) between internal subsystems of system $\mathcal{X}$ affect the viability of the system. This would uncover 'internal' semantic information which would be involved in internal self-maintenance processes, such as homeostasis.

6. Automatic identification of initial distributions, timescales and decompositions of interest

Our measures of semantic information depend on: (1) the decomposition of the world into the system $\mathcal{X}$ and the environment $\mathcal{Y}$; (2) the timescale τ; and (3) the initial distribution over joint states of the system and environment. The factors generally represent 'subjective' choices of the scientist, indicating for which systems, temporal scales, and initial conditions the scientist wishes to quantify semantic information.

However, it is also possible to select these factors in a more 'objective' manner, in particular by choosing decompositions, timescales, and initial distributions for which semantic information measures—such as the value of information or the amount of semantic information—are maximized.

For example, consider fixing a particular timescale τ and a particular decomposition into system/environment, and then identifying the initial distribution which maximizes the viability value of stored semantic information,

$$p^{\star}_{X_0,Y_0} \in \underset{q_{X_0,Y_0}}{\operatorname{argmax}}\ \Delta V^{\text{stored}}_{\text{tot}}(q_{X_0,Y_0}), \tag{6.1}$$

where we have made the dependence of ΔV_{tot} on the initial distribution explicit in equation (6.1), but left implicit its dependence on the timescale τ and the decomposition into $\mathcal{X}$ and $\mathcal{Y}$. Given the intrinsic dynamics of the system and environment, $p^{\star}_{X_0,Y_0}$ captures the initial distribution that the system is 'best fit for' in an informational sense, i.e. the distribution under which the system most benefits from having syntactic information about the environment. One can then define various other semantic information measures, such as the amount of semantic information and the semantic content of particular states, relative to $p^{\star}_{X_0,Y_0}$, rather than some exogenously specified initial distribution. For instance, the semantic content of some system state $x \in X$ under $p^{\star}_{X_0,Y_0}$ represents the conditional distribution over environmental states that, given the dynamics of system and environment, x is 'best fit to represent' in terms of maximizing viability value.

One can also maximize the value of information (or other measures) over timescales τ and system/environment decompositions of the world, so as to automatically detect subsystems and temporal scales that exhibit large amounts of semantic information. As mentioned in the Introduction, our work is conceptually inspired by work on autonomous agents, and our approach in fact suggests a possible formal and quantitative definition of **autonomous agency**: a physical system is an autonomous agent to the extent that it has a large measure of semantic information. From this point of view, finding timescales and system/environment decompositions that maximize measures of semantic information provides a way to automatically identify agents in the physical world (see also [141–144]). Exploring these possibilities, including which semantic information measures (value of information, the amount of semantic information, thermodynamic multiplier, etc.) are best for automatically identifying agents, remains for future work.

7. Conclusion and discussion

In this paper, we propose a definition of semantic information as the syntactic information between a physical system and its environment that is causally necessary for maintaining the system's existence. We consider two particular measures of semantic information: stored semantic information, which is based on the mutual information between system and environment at $t = 0$, and observed semantic information, which is based on the transfer entropy exchanged between system and environment over $t \in [0, \tau]$.

Our measures possess several features that have been proposed as desirable characteristics of any measure of semantic information in the philosophical literature [3,4,6]. Unlike syntactic information, semantic information should be able to be 'mistaken', i.e. to 'misrepresent' the world. This emerges naturally in our framework whenever information has a negative viability value (i.e. when the system uses information in a way that actually hurts its ability to maintain its own existence). Furthermore, a notion of semantic information between a system and environment should be fundamentally *asymmetrical* (unlike some measures of syntactic information, such as mutual information). For instance, a chemotactic bacterium swimming around a nutrient solution is presumed to have semantic information about its environment, but the environment is not expected to have semantic information about the bacterium. Our measures of semantic information are fundamentally asymmetrical—even when defined relative to a symmetric syntactic information measure like mutual information—because they are defined in terms of their contribution to the viability of the system, rather than the environment.

Our framework does not require the system of interest to be decomposed into separate degrees of freedom representing 'sensors' versus 'effectors' (or 'membrane' versus 'interior', 'body' versus 'brain', etc.). This is advantageous because such distinctions may be difficult or impossible to define for certain systems. Our framework also side-steps questions of what type of 'internal models' or 'internal representations', if any, are employed by the system. Instead, our definitions of semantic information, including the semantic content of particular states of the system, are grounded in the intrinsic dynamics of the system and environment.

As mentioned, we do not assume that the system of interest is an organism. At the same time, in cases where the system is, in fact, an organism (or an entire population of organisms) undergoing an evolutionary process, there are promising connections between our approach and information-theoretic ideas in theoretical biology. For instance, various ways of formalizing fitness-relevant information in biology [144–146] appear conceptually, and perhaps formally, related to our definitions of semantic information. Exploring such connections remains for future work.

Organisms are, of course, the prototypical self-maintaining systems, and will generally have high levels of both stored and observed semantic information. This suggests that our measures of semantic information may be useful as part of quantitative, formal definitions of life. In particular, we suggest that having high levels of semantic information is a necessary, though perhaps not sufficient, condition for any physical system to be alive.

Data accessibility. This article has no additional data.

Competing interests. We declare we have no competing interests.

Funding. This paper was made possible through grant no. FQXi-RFP-1622 from the FQXi foundation, and grant no. CHE-1648973 from the US National Science Foundation.

Acknowledgments. We thank Carlo Rovelli, Daniel Polani, Jacopo Grilli, Chris Kempes and Mike Price, along with Luis Bettencourt, Daniel Dennett, Peter Godrey-Smith, Chris Wood and other participants in the Santa Fe Institute's 'Meaning of Information' working group for helpful conversations. We also thank the anonymous reviewers for helpful suggestions. We thank the Santa Fe Institute for helping to support this research.

Appendix A. Relationship between entropy and probability of being in viability set

Imagine that $\mathcal{A} \subseteq X$ is some set of desirable states, which we call the *viability set*. Assume that $|\mathcal{A}| \ll |X|$. Here we show that entropy bounds the probability that $\mathcal{X}$ is in set $\mathcal{A}$ as

$$p(X \in \mathcal{A}) = \sum_{x \in \mathcal{A}} p(x) \lesssim \frac{\log |X| - S(p(X))}{\log |X| - \log |\mathcal{A}|}. \tag{A 1}$$

To demonstrate this, let $\mathbf{1}_{\mathcal{A}}(x)$ be the indicator function for set $\mathcal{A}$, so that $\mathbf{1}_{\mathcal{A}}(x)$ is equal to 1 when $x \in \mathcal{A}$, and 0 otherwise. Using the chain rule for entropy, we write

$$\begin{aligned} S(p(X)) &= S(p(X, \mathbf{1}_{\mathcal{A}}(X))) \\ &= S(p(X|\mathbf{1}_{\mathcal{A}}(X))) + S(\mathbf{1}_{\mathcal{A}}(X)) \\ &\leq S(p(X|\mathbf{1}_{\mathcal{A}}(X))) + 1. \end{aligned} \tag{A 2}$$

In the last line, we use the fact that the maximum entropy of a binary random variable, such as $\mathbf{1}_{\mathcal{A}}(X)$, is 1 bit.

We now rewrite the conditional entropy as

$$\begin{aligned} S(p(X|\mathbf{1}_{\mathcal{A}}(X))) &= p(X \in \mathcal{A}) \cdot S(X|X \in \mathcal{A}) + (1 - p(X \in \mathcal{A})) \\ &\quad \cdot S(X|X \notin \mathcal{A}) \\ &\leq p(X \in \mathcal{A}) \log |\mathcal{A}| + (1 - p(X \in \mathcal{A})) \log |X \backslash \mathcal{A}|, \end{aligned} \tag{A 3}$$

where we have used the fact that entropy of any distribution over a set of size n is upper bounded by $\log n$ (as achieved by the uniform distribution over that set). Combining with equation (A 2) gives

$$S(p(X)) \leq p(X \in \mathcal{A})(\log |\mathcal{A}| - \log |X \backslash \mathcal{A}|) + \log |X \backslash \mathcal{A}| + 1.$$

Rearranging gives

$$\begin{aligned} p(X \in \mathcal{A}) &\leq \frac{-S(p(X)) + \log |X \backslash \mathcal{A}| + 1}{\log |X \backslash \mathcal{A}| - \log |\mathcal{A}|} \\ &= 1 - \frac{S(p(X)) - \log |\mathcal{A}| - 1}{\log |X \backslash \mathcal{A}| - \log |\mathcal{A}|} \\ &\leq 1 - \frac{S(p(X)) - \log |\mathcal{A}| - 1}{\log |X| - \log |\mathcal{A}|} \\ &\approx 1 - \frac{S(p(X)) - \log |\mathcal{A}|}{\log |X| - \log |\mathcal{A}|} \\ &= \frac{\log |X| - S(p(X))}{\log |X| - \log |\mathcal{A}|}, \end{aligned}$$

where we have dropped the $1/(\log |X \backslash \mathcal{A}| - \log |\mathcal{A}|)$ term.

Thus, as entropy goes up, the probability concentrated within any small set goes down.

Appendix B. Model of food-seeking agent

In this appendix, we describe our model of a simple food-seeking system.

In this model, the state space of the environment $\mathcal{Y}$ consists of $Y = \{1..n\} \cup \{\emptyset\}$, representing the location of a single unit of food along 1 spatial dimension, or the possible lack of food ($\emptyset$). The state space of the agent (i.e. the system $\mathcal{X}$) consists of three separate degrees of freedom, indicated as $X = X^{\text{loc}} \times X^{\text{level}} \times X^{\text{target}}$. $X^{\text{loc}} = \{1..n\}$ represents the spatial location of the agent out of n possible locations. $X^{\text{level}} = \{0..l_{\max}\}$ represents the 'satiation level' of the agent, ranging from 'fully fed' ($l_{\max}$) to 'dead' (0). $X^{\text{target}} = \{1..n\} \cup \{\emptyset\}$ represents the agent's internal information about the location of food in the environment ($\emptyset$ corresponding to information that there is no food).

The dynamics are such that, as long as the agent is not 'dead' ($X^{\text{level}} \neq 0$), the agent moves towards X^{target}. If the agent reaches a location sufficiently close to the food ($|X^{\text{loc}} - Y| \leq 1$), the agent 'eats the food', meaning that satiation level of the agent is changed to $l_{\max}$. Otherwise, the satiation level drops by one during every timestep. The food stays in the same place unless it gets eaten, or unless it spontaneously degrades (goes to $\emptyset$) which happens with a small probability in each step. The agent never changes its target belief. All states are assigned free energy values, for which the dynamics obey local detailed balance.

Initially, the agent is located at the centre spatial location ($X_0^{\text{loc}} = \lfloor n/2 \rfloor$), the satiation level is maximal $X_0^{\text{level}} = l_{\max}$, the food location is uniformly distributed over $1..n$, and the agent has perfect information about the location of the food, $p(x_0^{\text{target}}|y_0) = \delta(x_0^{\text{target}}, y_0)$.

We assume that the state space of the agent corresponds to a set coarse-grained macrostates. Formally, we write this as $X = f(Z)$, where Z is a random variable indicating the microstate of $\mathcal{X}$ and f is a function that maps from microstates to macrostates. The entropy of any microstate distribution p_{Z_τ} can be written as

$$\begin{aligned} S(p_{Z_\tau}) &= S(p_{Z_\tau, X_\tau}) \\ &= S(p_{X_\tau}) + S(p_{Z_\tau|X_\tau}) \\ &= S(p_{X_\tau}) + \sum_{x_\tau} p(x_\tau) S(p_{Z_\tau|f(Z_\tau)=x_\tau}). \end{aligned}$$

We assume that within each macrostate, the microstate distribution relaxes instantly to some local equilibrium, so that each 'internal entropy' term $S(p_{Z|f(Z)=x_\tau})$ is constant, which

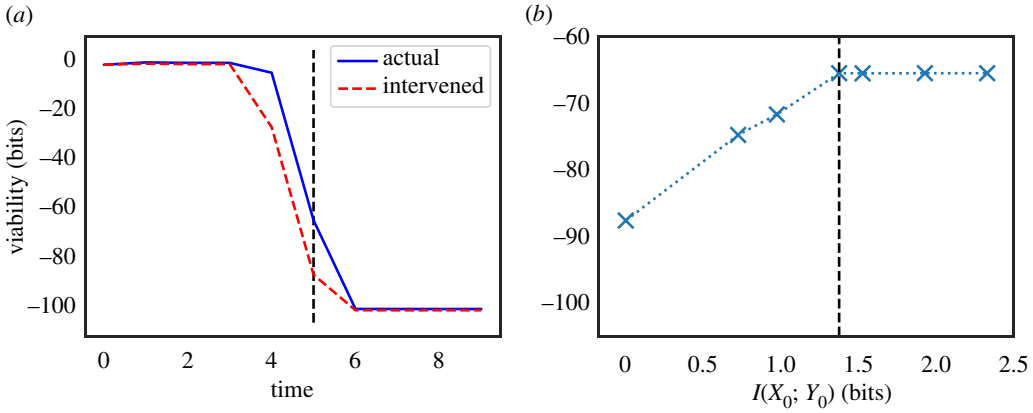

Figure 3. Illustration of our approach using a simple model of a food-seeking system. Under the actual distribution, the system has perfect knowledge of the location of food at $t = 0$. (*a*) We plot viability values over time under both the actual and (fully scrambled) intervened distributions. The vertical dashed line corresponds to our timescale of interest ($\tau = 5$ timesteps). (*b*) We plot the information/viability curve for $\tau = 5$ ($\times$'s are actual points on the curve, dashed line is interpolation). The vertical dashed line indicates the amount of stored semantic information. See text for details. (Online version in colour.)

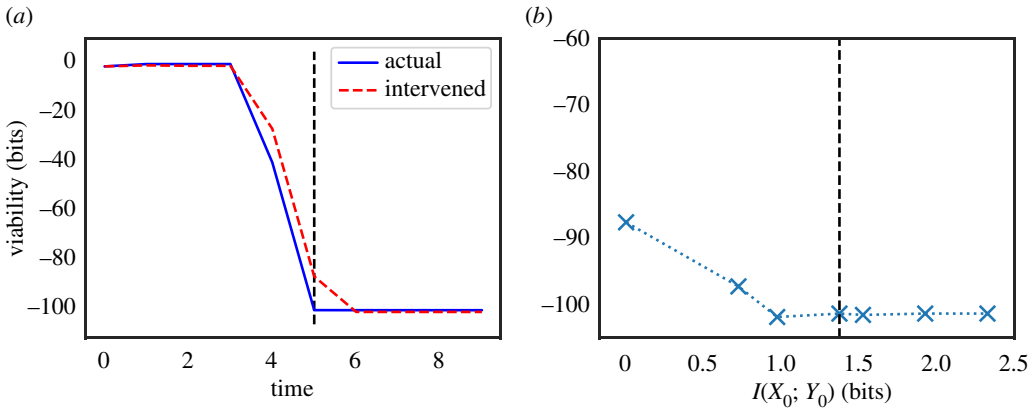

Figure 4. Illustration of our measure with a simple model of a system which moves *away* from where it believes food to be located. (*a*) We plot viability values over time under both the actual and (fully scrambled) intervened distributions. The vertical dashed line corresponds to our timescale of interest ($\tau = 5$ timesteps). (*b*) We plot the information/viability curve for $\tau = 5$ ($\times$'s are actual points on the curve, dashed line is interpolation). The vertical dashed line indicates the amount of stored semantic information. See text for details. (Online version in colour.)

we indicate as $S_{\text{int}}(x_\tau)$. Combining, we compute our negative entropy viability function as

$$V(p_{X_\tau}) = S(p_{X_\tau}) + \sum_{x_\tau} p(x_\tau) S_{\text{int}}(x_\tau).$$

In this particular model, we take the internal entropy of all macrostates to be 0, except for any macrostate which has $X^{\text{level}} = 0$ (i.e. the agent is 'dead'), in which case the internal entropy is S_{dead} bits. Essentially, this means that the system equilibrates instantly within the dead macrostate, and that the dead macrostate has a large internal entropy (i.e. there are many more ways of being dead than not).

To avoid having results that are sensitive to numerical errors, we 'smooth' the information/viability curve by rounding all viability and mutual information values to 5 decimal places.

Figure 3 shows the results for parameters $n = 5$, $l_{\max} = 5$, $S_{\text{dead}} = 100$ bits and timescale $\tau = 5$. The total amount of mutual information is $\log_2 5 \approx 2.32$ bits, while the total amount of semantic information is only ≈ 1.37 bits, which gives a semantic efficiency value of $\kappa_{\text{stored}} \approx 0.6$. This occurs because if the food is initially in locations $\{2, 3, 4\}$, the agent is close enough to eat it immediately, and knowing in which of these three locations the food is located does not affect viability. The viability value of information is $\Delta V_{\text{tot}}^{\text{stored}} \approx 22.1$ bits, giving a thermodynamic multiplier of $\kappa_{\text{stored}} \approx 9.5$. The model is also discussed in §5.1.4 in the main text.

We also analyse a different model, in which the agent's dynamics are such that it moves *away* from the target in each timestep, until it reaches the edges of the world ($X^{\text{loc}} = 1$ or $X^{\text{loc}} = n$) and stays there. The agent still dies if it does not eat food for some number of timesteps. As before, the agent begins initially with perfect information about the location of the food. In this case, information about the world actually hurts the agent's ability to maintain its own existence, leading to a negative viability value of information.

Figure 4 shows the results for this model, using the same parameter values as before ($n = 5$, $l_{\max} = 5$, $S_{\text{dead}} = 100$ bits and timescale $\tau = 5$). The total amount of mutual information is again $\log_2 5 \approx 2.32$ bits, and the total amount of semantic information is again ≈ 1.37 bits (if the food is initially in locations $\{2, 3, 4\}$, the system is close enough to eat it immediately, and knowing in which of these three locations the food is located does not affect viability). This gives a semantic efficiency value of $\kappa_{\text{stored}} \approx 0.6$. Unlike the food-seeking agent, the viability value of information in this case is $\Delta V_{\text{tot}}^{\text{stored}} \approx -13.7$ bits, giving a thermodynamic multiplier of $\kappa_{\text{stored}} \approx -5.9$.

A Python implementation of these models is available at https://github.com/artemyk/semantic_information/.

Endnotes

[1]Semantic information has also sometimes been called 'meaningful information' [25–28], 'relevant information' [19,20], 'functional information' [29,30] and 'pragmatic information' [9] in the literature.

[2]Much of our approach can also be used to quantify semantic information in any dynamical system, not just physical systems. For the purposes of this paper, however, we focus our attention on physical systems.

[3]Interestingly, the thermodynamic multiplier is related to an information-theoretic measure of efficiency of closed-loop control suggested in [62, eqn 54].

[4]We assume that the conditional distribution $p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$ is fully specified. This is always the case if the conditional distribution $p_{X_{t+1}|X_t,Y_t}$ is strictly positive for all x_t, y_t, x_{t+1}, since then $p(y_{t+1}|x_t, y_t, x_{t+1}) = p(x_{t+1}, y_{t+1}|x_t, y_t)/p(x_{t+1}|x_t, y_t)$. If $p_{X_{t+1}|X_t,Y_t}$ is not strictly positive, then $p_{Y_{t+1}|X_t,Y_t,X_{t+1}}$ has to be explicitly provided, e.g. by specifying the joint stochastic dynamics via an appropriate Bayesian network.

References

1. Dretske F. 1981 *Knowledge and the flow of information*. Cambridge, MA: MIT Press.
2. Shea N. 2007 Representation in the genome and in other inheritance systems. *Biol. Phil.* **22**, 313–331. (doi:10.1007/s10539-006-9046-6)
3. Godfrey-Smith P, Sterelny K. 2016 Biological information. In *The Stanford encyclopedia of philosophy* (ed. EN Zalta), summer 2016 edn. Metaphysics Research Lab, Stanford University.
4. Godfrey-Smith P. 2007 Information in biology. In *The Cambridge companion to the philosophy of biology* (eds D Hull, M Ruse). Cambridge Companions to Philosophy, pp. 103–119. Cambridge, UK: Cambridge University Press. (doi:10.1017/CCOL9780521851282.006)
5. Collier J. 2008 Information in biological systems. In *Handbook of philosophy of science. Volume 8: Philosophy of information*, pp. 763–787. Amsterdam, The Netherlands, Elsevier.
6. Sterelny K, Griffiths PE. 1999 *Sex and death: an introduction to philosophy of biology*. Chicago, IL: University of Chicago Press.
7. Millikan RG. 1984 *Language, thought, and other biological categories: new foundations for realism*. Cambridge, MA: MIT Press.
8. Jablonka E. 2002 Information: its interpretation, its inheritance, and its sharing. *Phil. Sci.* **69**, 578–605. (doi:10.1086/344621)
9. Weinberger ED. 2002 A theory of pragmatic information and its application to the quasi-species model of biological evolution. *Biosystems* **66**, 105–119. (doi:10.1016/S0303-2647(02)00038-2)
10. Barham J. 1996 A dynamical model of the meaning of information. *Biosystems* **38**, 235–241. (doi:10.1016/0303-2647(95)01596-5)
11. Ruiz-Mirazo K, Moreno A. 2004 Basic autonomy as a fundamental step in the synthesis of life. *Artif. Life* **10**, 235–259. (doi:10.1162/1064546041255584)
12. Moreno A, Etxeberria A. 2005 Agency in natural and artificial systems. *Artif. Life* **11**, 161–175. (doi:10.1162/1064546053278919)
13. Deacon TW. 2008 Shannon-Boltzmann-Darwin: redefining information (Part II). *Cogn. Semiotics* **2**, 169–196. (doi:10.1515/cogsem.2008.2.spring2008.169)
14. Bickhard MH. 2000 Autonomy, function, and representation. *Commun. Cogn. Artif. Intell.* **17**, 111–131.
15. Harnad S. 1990 The symbol grounding problem. *Phys. D Nonlinear Phenom.* **42**, 335–346. (doi:10.1016/0167-2789(90)90087-6)
16. Glenberg AM, Robertson DA. 2000 Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* **43**, 379–401. (doi:10.1006/jmla.2000.2714)
17. Steels L. 2003 Evolving grounded communication for robots. *Trends Cogn. Sci.* (*Regul. Ed.*) **7**, 308–312. (doi:10.1016/S1364-6613(03)00129-3)
18. Crutchfield JP. 1992 Semantics and thermodynamics. In *Santa Fe Institute studies in the sciences of complexity*, vol. 12, pp. 317–317.
19. Tishby N, Pereira FC, Bialek W. 2000 The information bottleneck method. (http://arxiv.org/abs/physics/0004057)
20. Polani D, Nehaniv CL, Martinetz T, Kim JT. 2006 Relevant information in optimized persistence vs. progeny strategies. In *Artificial Life X: Proc. of the Tenth Int. Conf. on the Simulation and Synthesis of Living Systems*. Cambridge, MA: MIT Press.
21. Vitányi PM. 2006 Meaningful information. *IEEE Trans. Inf. Theory* **52**, 4617–4626. (doi:10.1109/TIT.2006.881729)
22. Bar-Hillel Y, Carnap R. 1953 Semantic information. *Br. J. Phil. Sci.* **4**, 147–157. (doi:10.1093/bjps/IV.14.147)
23. Dennett DC. 1983 Intentional systems in cognitive ethology: the 'panglossian paradigm' defended. *Behav. Brain Sci.* **6**, 343–355. (doi:10.1017/S0140525X00016393)
24. Floridi L. 2005 Semantic conceptions of information. In *The Stanford encyclopedia of philosophy* (ed. EN Zalta), spring 2017 edn.
25. Nehaniv CL. 1999 Meaning for observers and agents. In *Proc. 1999 IEEE Int. Symp. on Intelligent Control/Intelligent Systems and Semiotics*, pp. 435–440. (doi:10.1109/ISIC.1999.796694)
26. Nehaniv CL, Polani D, Dautenhahn K, te Boekhorst R, Cañamero L. 2002 Meaningful information, sensor evolution, and the temporal horizon of embodied organisms. In *Proc. Artificial Life VIII*, pp. 345–349. Cambridge, MA: MIT Press.
27. Atlan H. 1987 Self creation of meaning. *Phys. Scr.* **36**, 563–576. (doi:10.1088/0031-8949/36/3/032)
28. Reading A. 2011 *Meaningful information*, pp. 9–15. Berlin, Germany: Springer.
29. Farnsworth KD, Nelson J, Gershenson C. 2013 Living is information processing: from molecules to global systems. *Acta Biotheor.* **61**, 203–222. (doi:10.1007/s10441-013-9179-3)
30. Sharov AA. 2010 Functional information: towards synthesis of biosemiotics and cybernetics. *Entropy* **12**, 1050–1070. (doi:10.3390/e12051050)
31. Wolpert D, Kolchinsky A. 2016 Observers as systems that acquire information to stay out of equilibrium. In *FQXi's 5th International Conference*. See https://www.youtube.com/watch?v = zVpSAjAe-tE.
32. Rovelli C. 2018 Meaning and intentionality=information+evolution. In *Wandering towards a goal* (eds A Aguirre, B Foster, Z Merali), pp. 17–27. Cham, Switzerland: Springer. (doi:10.1007/978-3-319-75726-1_3)
33. Seifert U. 2012 Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.* **75**, 126001. (doi:10.1088/0034-4885/75/12/126001)
34. Parrondo JM, Horowitz JM, Sagawa T. 2015 Thermodynamics of information. *Nat. Phys.* **11**, 131–139. (doi:10.1038/nphys3230)
35. Shannon CE. 1948 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656.
36. Cover TM, Thomas JA. 2012 *Elements of information theory*. New York, NY: John Wiley & Sons.
37. von Neuman J, Morgenstern O. 1944 *Theory of games and economics behavior*. Princeton, NJ: Princeton University Press.
38. Fudenberg D, Tirole J. 1991 *Game theory*. Cambridge, MA: MIT Press.
39. Polani D, Martinetz T, Kim J. 2001 An information-theoretic approach for the quantification of relevance. In *Advances in artificial life*, pp. 704–713. Berlin, Germany: Springer.
40. Gould JP. 1974 Risk, stochastic preference, and the value of information. *J. Econ. Theory* **8**, 64–84. (doi:10.1016/0022-0531(74)90006-4)
41. Hess J. 1982 Risk and the gain from information. *J. Econ. Theory* **27**, 231–238. (doi:10.1016/0022-0531(82)90026-6)
42. Harremoes P, Tishby N. 2007 The information bottleneck revisited or how to choose a good distortion measure. In *2007 IEEE Int. Symp. on Information Theory, Nice, France, 24–29 June 2007*, pp. 566–570. (doi:10.1109/ISIT.2007.4557285)
43. DeGroot MH. 1962 Uncertainty, information, and sequential experiments. *Ann. Math. Stat.* **33**, 404–419. (doi:10.1214/aoms/1177704567)
44. Jiao J, Courtade T, Venkat K, Weissman T. 2015 Justification of logarithmic loss via the benefit of side information. *IEEE Trans. Inf. Theory* **61**, 5357–5365. (doi:10.1109/TIT.2015.2462848)
45. Schlosser G. 1998 Self-re-production and functionality. *Synthese* **116**, 303–354. (doi:10.1023/A:1005073307193)

46. Mossio M, Saborido C, Moreno A. 2009 An organizational account of biological functions. *Br. J. Phil. Sci.* **60**, 813–841. (doi:10.1093/bjps/axp036)
47. Maturana HR, Varela FJ. 1991 *Autopoiesis and cognition: the realization of the living*, vol. 42. Berlin, Germany: Springer Science & Business Media.
48. Thompson E, Stapleton M. 2008 Making sense of sense-making: reflections on enactive and extended mind theories. *Topoi* **28**, 23–30. (doi:10.1007/s11245-008-9043-2)
49. Froese T, Ziemke T. 2009 Enactive artificial intelligence: investigating the systemic organization of life and mind. *Artif. Intell.* **173**, 466–500. (doi:10.1016/j.artint.2008.12.001)
50. Raymond RC. 1950 Communication, entropy, and life. *Am. Sci.* **38**, 273–278.
51. Collier J. 1990 Intrinsic information. In *Information, language and cognition* (ed. PP Hanson), pp. 390–409. Vancouver, Canada: University of British Columbia Press.
52. Collier JD, Hooker CA. 1999 Complexly organised dynamical systems. *Open Syst. Inf. Dyn.* **6**, 241–302. (doi:10.1023/A:1009662321079)
53. Friston KJ, Stephan KE. 2007 Free-energy and the brain. *Synthese* **159**, 417–458. (doi:10.1007/s11229-007-9237-y)
54. Kauffman S. 2003 Molecular autonomous agents. *Phil. Trans. R. Soc. Lond. A* **361**, 1089–1099. (doi:10.1098/rsta.2003.1186)
55. Maes P. 1990 *Designing autonomous agents: theory and practice from biology to engineering and back*. Cambridge, MA: MIT Press.
56. Melhuish C, Ieropoulos I, Greenman J, Horsfield I. 2006 Energetically autonomous robots: food for thought. *Auton. Robots* **21**, 187–198. (doi:10.1007/s10514-006-6574-5)
57. McGregor S, Virgo N. 2009 Life and its close relatives. In *European Conf. on Artificial Life*, pp. 230–237. Berlin, Germany: Springer.
58. Barandiaran XE, Di Paolo E, Rohde M. 2009 Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* **17**, 367–386. (doi:10.1177/1059712309343819)
59. Chandrasekhar S. 1961 *Hydrodynamic and hydromagnetic stability*. Oxford, UK: Clarendon Press.
60. Thompson E. 2007 *Mind in life: biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.
61. Touchette H, Lloyd S. 2000 Information-theoretic limits of control. *Phys. Rev. Lett.* **84**, 1156–1159. (doi:10.1103/PhysRevLett.84.1156)
62. Touchette H, Lloyd S. 2004 Information-theoretic approach to the study of control systems. *Phys. A Stat. Mech. Appl.* **331**, 140–172. (doi:10.1016/j.physa.2003.09.007)
63. Cao FJ, Feito M. 2009 Thermodynamics of feedback controlled systems. *Phys. Rev. E* **79**, 041118. (doi:10.1103/PhysRevE.79.041118)
64. Ashby WR. 1960 *Design for a brain*. London, UK: Chapman & Hall.
65. Beer RD. 1996 A dynamical systems perspective on agent-environment interaction. *Comput. Theories Interact. Agency* **72**, 173–215. (doi:10.1016/0004-3702(94)00005-l)
66. Di Paolo EA. 2005 Autopoiesis, adaptivity, teleology, agency. *Phenomenol. Cogn. Sci.* **4**, 429–452. (doi:10.1007/s11097-005-9002-y)
67. Agmon E, Gates AJ, Churavy V, Beer RD. 2016 Exploring the space of viable configurations in a model of metabolism-boundary co-construction. *Artif. Life* **22**, 153–171. (doi:10.1162/artl_a_00196)
68. Esposito M, Van den Broeck C. 2011 Second law and Landauer principle far from equilibrium. *Europhys. Lett.* **95**, 40004. (doi:10.1209/0295-5075/95/40004)
69. Maroney O. 2009 Generalizing Landauer's principle. *Phys. Rev. E* **79**, 031105. (doi:10.1103/PhysRevE.79.031105)
70. Wolpert DH. 2016 Extending Landauer's bound from bit erasure to arbitrary computation. (http://arxiv.org/abs/1508.05319)
71. Horowitz JM, Zhou K, England JL. 2017 Minimum energetic cost to maintain a target nonequilibrium state. *Phys. Rev. E* **95**, 042102. (doi:10.1103/physreve.95.042102)
72. Pearl J. 2000 *Causality: models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
73. Ay N, Polani D. 2008 Information flows in causal networks. *Adv. Complex Syst.* **11**, 17–41. (doi:10.1142/S0219525908001465)
74. Blackwell D. 1953 Equivalent comparisons of experiments. *Ann. Math. Stat.* **24**, 265–272. (doi:10.1214/aoms/1177729032)
75. Lindley DV. 1956 On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005. (doi:10.1214/aoms/1177728069)
76. Shannon CE. 1958 A note on a partial ordering for communication channels. *Inf. Control* **1**, 390–397. (doi:10.1016/S0019-9958(58)90239-0)
77. Rauh J, Banerjee PK, Olbrich E, Jost J, Bertschinger N, Wolpert D. 2017 Coarse-graining and the Blackwell order. *Entropy* **19**, 527. (doi:10.3390/e19100527)
78. Nasser R. 2017 On the input-degradedness and input-equivalence between channels. (http://arxiv.org/abs/1702.00727)
79. Nasser R. 2017 A characterization of the Shannon ordering of communication channels. In *2017 IEEE Int. Symp. on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017*, pp. 2448–2452. (doi:10.1109/ISIT.2017.8006969)
80. Still S. 2017 Thermodynamic cost and benefit of data representations. (http://arxiv.org/abs/1705.00612)
81. Schreiber T. 2000 Measuring information transfer. *Phys. Rev. Lett.* **85**, 461–464. (doi:10.1103/PhysRevLett.85.461)
82. Schrodinger E. 1944 *What is life?* Cambridge, UK: Cambridge University Press.
83. Brillouin L. 1949 Life, thermodynamics, and cybernetics. *Am. Sci.* **37**, 554–568.
84. Bauer E. 1920 Die definition des lebewesens auf grund seiner thermodynamischen eigenschaften und die daraus folgenden biologischen grundprinzipien. *Naturwissenschaften* **8**, 338–340. (doi:10.1007/BF02448266)
85. Elek G, Müller M. 2012 The living matter according to Ervin Bauer (1890–1938). *Acta Physiol. Hung.* **100**, 124–132. (doi:10.1556/APhysiol.99.2012.006)
86. Morowitz HJ. 1955 Some order-disorder considerations in living systems. *Bull. Math. Biophys.* **17**, 81–86. (doi:10.1007/bf02477985)
87. Bonchev D, Kamenski D, Kamenska V. 1976 Symmetry and information content of chemical structures. *Bull. Math. Biol.* **38**, 119–133. (doi:10.1016/s0092-8240(76)80029-8)
88. Davies PCW, Rieper E, Tuszynski JA. 2013 Self-organization and entropy reduction in a living cell. *Biosystems* **111**, 1–10. (doi:10.1016/j.biosystems.2012.10.005)
89. Kempes CP, Wolpert D, Cohen Z, Pérez-Mercader J. 2017 The thermodynamic efficiency of computations made in cells across the range of life. *Phil. Trans. R. Soc. A* **375**, 20160343. (doi:10.1098/rsta.2016.0343)
90. Corning PA, Kline SJ. 1998 Thermodynamics, information and life revisited, part 1: 'to be or entrophy'. *Syst. Res. Behav. Sci.* **15**, 273–294.
91. Corning PA, Kline SJ. 1998 Thermodynamics, information and life revisited, part II: 'thermoeconomics' and 'control information'. *Syst. Res. Behav. Sci.* **15**, 453–482. (doi:10.1002/(sici)1099-1743(199811/12)15:6<453::aid-sres201>3.0.co;2-u)
92. Corning PA. 2001 'Control information': the missing element in Norbert Wiener's cybernetic paradigm? *Kybernetes* **30**, 1272–1288. (doi:10.1108/EUM0000000006552)
93. Ben Jacob E, Shapira Y, Tauber AI. 2006 Seeking the foundations of cognition in bacteria: from Schrodinger's negative entropy to latent information. *Phys. A Stat. Mech. Appl.* **359**, 495–524. (doi:10.1016/j.physa.2005.05.096)
94. Polani D. 2009 Information: currency of life? *HFSP J.* **3**, 307–316.
95. Deacon TW. 2010 What is missing from theories of information? In *Information and the nature of reality: from physics to metaphysics* (eds P Davies, NH Gregersen), pp. 186–216. Cambridge, UK: Cambridge University Press.
96. Krakauer DC. 2011 Darwinian demons, evolutionary complexity, and information maximization. *Chaos Int. J. Nonlinear Sci.* **21**, 037110. (doi:10.1063/1.3643064)
97. Turgut S. 2009 Relations between entropies produced in nondeterministic thermodynamic processes. *Phys. Rev. E* **79**, 041102. (doi:10.1103/PhysRevE.79.041102)
98. Sagawa T. 2014 Thermodynamic and logical reversibilities revisited. *J. Stat. Mech. Theory Exp.* **2014**, P03025. (doi:10.1088/1742-5468/2014/03/P03025)
99. Sagawa T, Ueda M. 2008 Second law of thermodynamics with discrete quantum feedback control. *Phys. Rev. Lett.* **100**, 080403. (doi:10.1103/PhysRevLett.100.080403)

100. Sagawa T, Ueda M. 2009 Minimal energy cost for thermodynamic information processing: measurement and information erasure. *Phys. Rev. Lett.* **102**, 250602. (doi:10.1103/PhysRevLett.102.250602)
101. Sagawa T, Ueda M. 2012 Nonequilibrium thermodynamics of feedback control. *Phys. Rev. E* **85**, 021104. (doi:10.1103/PhysRevE.85.021104)
102. Prokopenko M, Lizier JT, Price DC. 2013 On thermodynamic interpretation of transfer entropy. *Entropy* **15**, 524–543. (doi:10.3390/e15020524)
103. Horowitz JM, Sandberg H. 2014 Second-law-like inequalities with information and their interpretations. *New J. Phys.* **16**, 125007. (doi:10.1088/1367-2630/16/12/125007)
104. Cafaro C, Ali SA, Giffin A. 2016 Thermodynamic aspects of information transfer in complex dynamical systems. *Phys. Rev. E* **93**, 022114. (doi:10.1103/PhysRevE.93.022114)
105. Spinney RE, Lizier JT, Prokopenko M. 2016 Transfer entropy in physical systems and the arrow of time. *Phys. Rev. E* **94**, 022135. (doi:10.1103/PhysRevE.94.022135)
106. Spinney RE, Prokopenko M, Lizier JT. 2017 Transfer entropy in continuous time, with applications to jump and neural spiking processes. *Phys. Rev. E* **95**, 032319. (doi:10.1103/PhysRevE.95.032319)
107. Ito S, Sagawa T. 2013 Information thermodynamics on causal networks. *Phys. Rev. Lett.* **111**, 180603. (doi:10.1103/PhysRevLett.111.180603)
108. Horowitz JM, Esposito M. 2014 Thermodynamics with continuous information flow. *Phys. Rev. X* **4**, 031015. (doi:10.1103/physrevx.4.031015)
109. Horowitz JM. 2015 Multipartite information flow for multiple Maxwell demons. *J. Stat. Mech. Theory Exp.* **2015**, P03006. (doi:10.1088/1742-5468/2015/03/P03006)
110. Hartich D, Barato AC, Seifert U. 2016 Sensory capacity: an information theoretical measure of the performance of a sensor. *Phys. Rev. E* **93**, 022116. (doi:10.1103/PhysRevE.93.022116)
111. Allahverdyan AE, Janzing D, Mahler G. 2009 Thermodynamic efficiency of information and heat flow. *J. Stat. Mech. Theory Exp.* **2009**, P09011. (doi:10.1088/issn.1742-5468)
112. Sagawa T, Ueda M. 2010 Generalized Jarzynski equality under nonequilibrium feedback control. *Phys. Rev. Lett.* **104**, 090602. (doi:10.1103/PhysRevLett.104.090602)
113. Esposito M, Schaller G. 2012 Stochastic thermodynamics for 'Maxwell demon' feedbacks. *Europhys. Lett.* **99**, 30003. (doi:10.1209/0295-5075/99/30003)
114. Horowitz JM, Parrondo JM. 2011 Thermodynamic reversibility in feedback processes. *Europhys. Lett.* **95**, 10005. (doi:10.1209/0295-5075/95/10005)
115. Horowitz JM, Vaikuntanathan S. 2010 Nonequilibrium detailed fluctuation theorem for repeated discrete feedback. *Phys. Rev. E* **82**, 061120. (doi:10.1103/PhysRevE.82.061120)
116. Munakata T, Rosinberg M. 2012 Entropy production and fluctuation theorems under feedback control: the molecular refrigerator model revisited. *J. Stat. Mech. Theory Exp.* **2012**, P05010. (doi:10.1088/1742-5468/2012/05/p05010)
117. Munakata T, Rosinberg M. 2013 Feedback cooling, measurement errors, and entropy production. *J. Stat. Mech. Theory Exp.* **2013**, P06014. (doi:10.1088/1742-5468/2013/06/P06014)
118. Ponmurugan M. 2010 Generalized detailed fluctuation theorem under nonequilibrium feedback control. *Phys. Rev. E* **82**, 031129. (doi:10.1103/PhysRevE.82.031129)
119. Kim KH, Qian H. 2007 Fluctuation theorems for a molecular refrigerator. *Phys. Rev. E* **75**, 022102. (doi:10.1103/PhysRevE.75.022102)
120. Mandal D, Quan H, Jarzynski C. 2013 Maxwell's refrigerator: an exactly solvable model. *Phys. Rev. Lett.* **111**, 030602. (doi:10.1103/PhysRevLett.111.030602)
121. Barato AC, Seifert U. 2013 An autonomous and reversible Maxwell's demon. *Europhys. Lett.* **101**, 60001. (doi:10.1209/0295-5075/101/60001)
122. Koski JV, Maisi VF, Sagawa T, Pekola JP. 2014 Experimental observation of the role of mutual information in the nonequilibrium dynamics of a Maxwell demon. *Phys. Rev. Lett.* **113**, 030601. (doi:10.1103/PhysRevLett.113.030601)
123. Mandal D, Jarzynski C. 2012 Work and information processing in a solvable model of Maxwell's demon. *Proc. Natl Acad. Sci. USA* **109**, 11 641–11 645. (doi:10.1073/pnas.1204263109)
124. Gokler C, Kolchinsky A, Liu Z-W, Marvian I, Shor P, Shtanko O, Thompson K, Wolpert D, Lloyd S. 2017 When is a bit worth much more than kT ln2? (https://arxiv.org/abs/1705.09598)
125. Wehrl A. 1978 General properties of entropy. *Rev. Mod. Phys.* **50**, 221–260. (doi:10.1103/RevModPhys.50.221)
126. Van Kampen NG. 1992 *Stochastic processes in physics and chemistry*, vol. 1. Amsterdam, The Netherlands: Elsevier.
127. Egbert M, Barandiaran X. 2011 Quantifying normative behaviour and precariousness in adaptive agency. In *ECAL 2011. 11th European Conf. on Artificial Life*. Cambridge, MA: MIT Press.
128. Agmon E, Gates AJ, Beer RD. 2016 The structure of ontogenies in a model protocell. *Artif. Life* **22**, 499–517. (doi:10.1162/ARTL_a_00215)
129. Schlögl F. 1985 Thermodynamic metric and stochastic measures. *Zeitschrift für Physik B Condensed Matter* **59**, 449–454. (doi:10.1007/BF01328857)
130. Schneider ED, Kay JJ. 1994 Life as a manifestation of the second law of thermodynamics. *Math. Comput. Model.* **19**, 25–48. (doi:10.1016/0895-7177(94)90188-0)
131. Esposito M, Van den Broeck C. 2010 Three faces of the second law. I. Master equation formulation. *Phys. Rev. E* **82**, 011143. (doi:10.1103/PhysRevE.82.011143)
132. Van den Broeck C, Esposito M. 2010 Three faces of the second law. II. Fokker-Planck formulation. *Phys. Rev. E* **82**, 011144. (doi:10.1103/PhysRevE.82.011144)
133. Zia R, Schmittmann B. 2007 Probability currents as principal characteristics in the statistical mechanics of non-equilibrium steady states. *J. Stat. Mech. Theory Exp.* **2007**, P07012. (doi:10.1088/1742-5468/2007/07/p07012)
134. Platini T. 2011 Measure of the violation of the detailed balance criterion: a possible definition of a 'distance' from equilibrium. *Phys. Rev. E* **83**, 011119. (doi:10.1103/PhysRevE.83.011119)
135. Baiesi M, Maes C. 2018 Life efficiency does not always increase with the dissipation rate. *J. Phys. Commun.* **2**, 045017. (doi:10.1088/2399-6528/aab654)
136. Bouma G. 2009 Normalized (pointwise) mutual information in collocation extraction. In *Proc. Biennial GSCL Conf., From Form to Meaning, Potsdam, Germany, 30 September–2 October 2009*, pp. 31–40.
137. DeWeese MR, Meister M. 1999 How to measure the information gained from one symbol. *Net. Comput. Neural Syst.* **10**, 325–340. (doi:10.1088/0954-898X_10_4_303)
138. Lizier JT, Prokopenko M, Zomaya AY. 2008 Local information transfer as a spatiotemporal filter for complex systems. *Phys. Rev. E* **77**, 026110. (doi:10.1103/PhysRevE.77.026110)
139. Lizier JT. 2014 Measuring the dynamics of information processing on a local scale in time and space. In *Directed information measures in neuroscience*, pp. 161–193. Berlin, Germany: Springer.
140. Doyle FJ, Huyett LM, Lee JB, Zisser HC, Dassau E. 2014 Closed-loop artificial pancreas systems: engineering the algorithms. *Diabetes Care* **37**, 1191–1197. (doi:10.2337/dc13-2108)
141. Polani D, Ikegami T, Biehl M. 2016 Towards information based spatiotemporal patterns as a foundation for agent representation in dynamical systems. In *Proc. Artificial Life Conf. 2016*, pp. 722–729. Cambridge, MA: MIT Press.
142. Biehl M, Polani D. 2017 Action and perception for spatiotemporal patterns. In *Proc. European Conf. on Artificial Life*, vol. 14, pp. 68–75. Cambridge, MA: MIT Press.
143. Balduzzi D *et al.* 2011 Detecting emergent processes in cellular automata with excess information. In *Proc. 11th European Conf. on the Synthesis and Simulation of Living Systems*, pp. 55–62. Cambridge, MA: MIT Press.
144. Krakauer D, Bertschinger N, Olbrich E, Ay N, Flack JC. 2014 The information theory of individuality. (http://arxiv.org/abs/1412.2447)
145. Adami C. 2004 Information theory in molecular biology. *Phys. Life Rev.* **1**, 3–22. (doi:10.1016/j.plrev.2004.01.002)
146. Donaldson-Matasci MC, Bergstrom CT, Lachmann M. 2010 The fitness value of information. *Oikos* **119**, 219–230. (doi:10.1111/j.1600-0706.2009.17781.x)