Agentic Information Theory: Ergodicity and Intrinsic Semantics of Information Processes

James P. Crutchfield[1,*] and Alexandra Jurgens[2,†]

[1] *Complexity Sciences Center and Physics Department University of California at Davis, One Shields Avenue, Davis, CA 95616*

[2] *GEOSTAT Team, INRIA – Bordeaux Sud Ouest 33405 Talence Cedex, France*

(Dated: July 31, 2025)

We develop information theory for the temporal behavior of memoryful agents moving through complex—structured, stochastic—environments. We introduce and explore information processes—stochastic processes produced by cognitive agents in real-time as they interact with and interpret incoming stimuli. We provide basic results on the ergodicity and semantics of the resulting time series of Shannon information measures that monitor an agent's adapting view of uncertainty and structural correlation in its environment.

Keywords: stochastic process, Shannon information measures, stationarity, ergodicity, entropy rate, statistical complexity, excess entropy

CONTENTS

* chaos@ucdavis.edu; http://csc.ucdavis.edu/~chaos/
† alexandra.jurgens@inria.fr; http://csc.ucdavis.edu/~ajurgens/

I. INTRODUCTION

Contemporary statistical mechanics [1–3] has made considerable contributions to machine learning [4–7], from formal models and algorithms (e.g., reinforcement learning [8, 9]) to implementing thermodynamic systems for learning tasks [10, 11]. This cross-fertilization reflects broader recent concerns, though, about emergent organization in active matter [12], collective intelligence [13], embodiments of biological intelligence [14], and distributed artificial intelligence. Whether groups of molecules, animals, or robots, these systems consist of many interacting components—often referred to generically as "agents" in the complex adaptive systems and machine learning literatures [8, 15–18]. These agents glean information from their environment and use it to take actions, that in turn, affect the environment. One notable phenomenon is that these collections can organize into temporal and spatial behaviors beyond those of the individual agents. Today, this kind of emergent collective organization is becoming a concern as machine learning moves ever closer to massive deployments of adaptive agents on globe-spanning communication networks.

Characteristically, though, statistical mechanical approaches to collective thermodynamic functioning and information processing focus on asymptotic quantities—such as, say, correlation functions, time- or ensemble-average rates of work extraction, dissipation, and information storage. As such, they give only the barest insight into the actual, online, and time-dependent operation of an agent's internal mechanisms—the underlying information heat engines [19–23] that support relevant behavior and functioning.

The following addresses informational aspects of such real-time adaptive agents, ultimately referring to the *information processes* that reflect how a stochastic dynamical system acquires, stores, and internally processes information. In effect, these real-time adaptive agents directly manipulate information available in observations. This view begins to move away from indirect appeals to frequentist or Bayesian probability for defining and estimating information, instead to directly operating on inputs that are innately informative. Fundamental to this is prediction. This informational approach, though, entails several computational challenges. Fortunately, Ref. [24] gives explicit and efficient methods for calculating the underlying informational measures from a stochastic process' optimal representation—the ϵ-machine. And so, the conceit in the following is that a cognitive agent implements computational mechanics to learn internal models of its environment.

A. Challenge

The following lays out the basic concept of information processes—stochastic processes that capture what a cognitive agent observes and measures—informationally—from a stochastic process generated by an environment. It reviews prerequisites from stochastic processes, ergodic theory, and information theory.

To emphasize, such agents behave as more than memoryless particles or memoryless input-output mappings. They process information. In this literal sense the agents are *cognitive*—as they observe their environment they, explicitly or implicitly, make various interpretations of stimuli—apparent correlation and uncertainty and available free energy, for example. The latter are expressed as kinds of information estimated in real-time as an agent adapts to the time series of incoming stimuli and behaves accordingly.

To understand the agent's moment-by-moment operation we track various quantities estimated during the agent's interactions with its environment. Specifically, we monitor the interactions via Shannon's information atoms [25]. In this setting, then, an agent is a transducer that maps from a given process (the environment) to one or several observational stochastic processes—information processes comprised of time series of instantaneous entropy rate (apparent randomness), bound information (apparent information storage), elusive information (apparent hidden information), and so on.

We show that if the environment being observed is stationary then the resulting information processes developed by the agent are stationary, under certain conditions. These conditions include observing the environment through a sliding-window of finite duration or through a function—a statistic—with limited-range memory. If so, then the agent's interpretations of the environment's behavior are statistically well-behaved. We also address subtleties of nonstationarity that arise at early times through transients during agent-environment synchronization [26] and through the agent's employing incorrect internal models. We give companion results on the ergodicity of such information processes. We then turn to discuss how information processes convey meaning to an agent with an internal model (correct or misleading) of its environment.

B. Overview

The immediately following section provides background, by drawing parallels and even motivations from prior results on complex adaptive systems. It then reviews elementary information theory and Shannon information measures. These are then adapted to the online temporal setting of a cognitive agent interacting with its environment—the stochastic process it observes. Cognitive agents are introduced using the structural theory of information in stochastic processes—computational mechanics [27].

In short, a cognitive agent builds a minimal optimal predictor—an ϵ-machine—of its environment. A key point

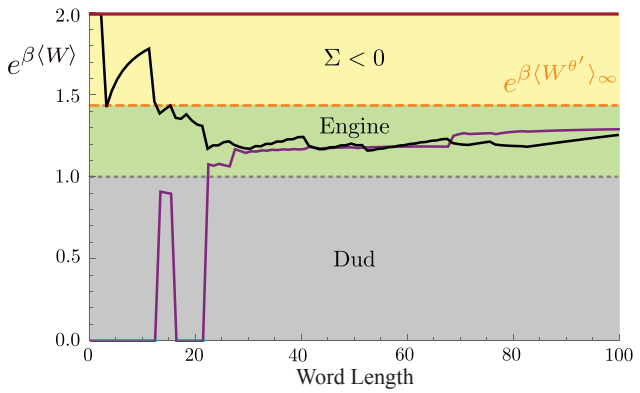

FIG. 1. Monitoring online thermodynamic performance via the work rate $\exp\beta\langle W^{\theta_\ell^{\max}}(y_{0:\ell})\rangle/\ell$ during learning: The information source is a length $\ell = 100$ sample output of a 5-state hidden Markov model (ϵ-machine); extracted from Fig. 6 in Ref. [28] (used with permission). The work rate is estimated using 2 memory states. The resulting work-rate monitoring processes are complicated and dynamic. Indeed, the engine jumps between different functionings as a "Dud" (gray)—unable to harvest work; an "Engine" (green)—doing so successfully; and violating the Second Law (yellow)—negative entropy fluctuation. The gray dashed line at 1.0 corresponds to zero work production: $\beta\langle W^{\theta}\rangle_\infty = 0$. The orange dashed line for work production corresponds to correctly estimating the true model $\beta\langle W^{\theta'}\rangle_\infty$. The solid red line at $\beta\langle W^{\theta}\rangle_\infty = \ln 2$ is the maximum work that can be harvested per bit from a single sequence.

is that agent behavior and its interpreting the environment are themselves stochastic processes. Taken altogether, this then gives the foundation for how an agent attributes meaning to the results of interacting with its environment. In short, there are two levels of semantics, one focused on the agent's uncertainties in predicting the environment, the other focused on the agent's informational interpretations of correlation and structure.

In this way, we develop methods and quantities to monitor how an agent interprets its interactions with a structured, stochastic environment. Here, we assume that the environment is stationary—its structure and stochasticity are time independent—and that the agent begins interacting by employing a given fixed model of the environment's stochasticity and structure—its internal model is time-independent. The agent only observes—it takes no actions on the environment. For now, these are the main simplifications. We then outline how the agent can use its internal model to interpret the environment's behavior moment-by-moment. One goal is to show that the resulting information processes are statistically well-behaved and so can form the basis for further—say, downstream—estimation and interpretation. Sequels relax these assumptions.

II. BACKGROUND

Before delving into technicalities it will be helpful later on to have in mind a range of example agent-environment systems—systems in which (i) extracting information from incoming signals and (ii) interpreting that information play a key role in agent behavior and collective functioning.

A. Collective Information Processes

A paradigmatic agent-environment system is that observed in a collective of flocking animals [29]. The social behaviors of fish shoaling and schooling and starling murmurations come immediately to mind [30, 31]. The animals are the agents that take in information from the spatial surroundings and, importantly, the nearby animals. From this they determine their motion. The result can be the emergence of stunningly, captivating complex coordinated spatio-temporal patterns.

One of the puzzles is how this complexity emerges from the given local, but distributed equations of motion. Using time-dependent, time-delayed mutual information Ref. [13] showed how a leader-follower relationship between animals emerges, eventually leading to a single individual guiding the flock's collective motion. In this, we see a real-world example of agents (i) interacting via an environment that (ii) effectively use one or several information processes to control local behaviors from which collective organization emerges.

B. Information Embedded in Thermodynamic Systems

While a primary concern here, information is not the only narrative framing for cognitive agents. For example, analogous *thermodynamic processes* describe an agent's creation and use of physical resources. Of late, Maxwell's demon is the oft-quoted example [32]. In this, agents are viewed as information heat engines. That is, there are other closely-related stochastic processes of interest. In particular, if we are interested in an agent's adaptive physical functioning, then a number of thermodynamic-resource processes capture engine performance.

The processes generated by Ref. [33]'s quantum machines and Ref. [34] self-correcting information engines come to mind. To be more concrete about such processes, however, consider a system that learns and then validates an information heat engine through environment interactions. Figure 1, excerpted from Ref. [28], shows the time evolution of an information engine's work rate $W^{\theta^{\max}}$. An exponential average work rate is plotted for both learning (purple line) and validation (black line) versus time series length ℓ. The point is that such agentic signals track

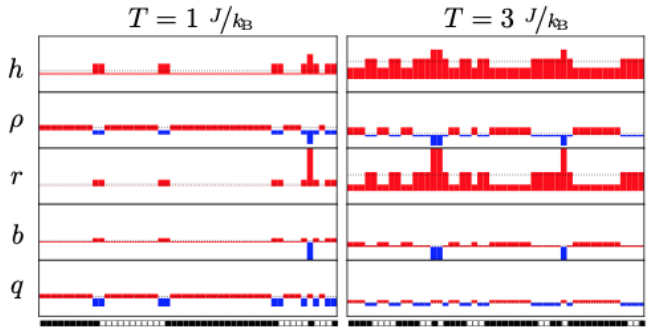

FIG. 2. Information processes in spin configurations: Motif entropy-component analysis of the 1D Ising model at two temperatures T, where h_μ is the local entropy density, ρ_μ the anticipated information, r_μ the ephemeral information, b_μ the bound information, and q_μ the enigmatic information. (J is the nearest-neighbor spin coupling strength and k_B is Boltzmann's constant.) A segment of a spin configuration with up spins (white cells) and down spins (black cells) shown in the bottom row. (Figure 6, Ref. [36], used with permission.)

the agent's evolution during learning and its functioning across time. They are stochastic processes—functions of the engine's internal, on-going adaptation and operation. And, they are signals in the design and application of information engines that are essential to monitor.

Relying on such signals, though, to monitor learning or online performance, one needs to know whether or not they are statistically well-behaved processes—signals that can be used to quantitatively (convergently) characterize learning and thermodynamic efficiency. Are they stationary and ergodic?

Reference [35] illustrates a higher-level of stochastic variation in an agent's very thermodynamic functioning. Is it acting as an engine extracting work from a heat reservoir or as an information eraser that correlates information? The several examples here illustrate that there are many circumstances when time-local instantaneous statistics and their operational meaning are key to monitoring adaptive physical behaviors.

C. Information Processes in Biology

Contrasted to physical systems for which information often simply stands in for probabilistic properties, information plays an essential role in the functioning of life processes [37]. Examples abound. For example, Ref. [38] argues that bacteria are environmental prediction engines—their reproduction and survival require them to take actions based the information that they extract and store from their surroundings, moment by moment. More to the point, it identifies *instantaneous predictive information* as a key signal—information shared between an organism's present phenotype and future environment

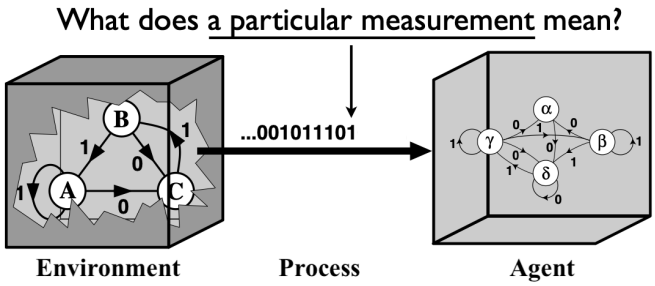

FIG. 3. Measurement semantics: The channel consists of the environment (left) that generates a process (middle)—the result of a series of observations and interpretations made by an agent (right).

states. And, it demonstrates that optimal epigenetic markers are minimal sufficient statistics for evolutionary prediction—the causal states introduced shortly. Bacteria in this setting are agents and their surroundings their environment.

Reference [39] develops a similar analysis of molecular sensors. In particular, it calls out the role of information gleaned by molecular sensors viewed as Markovian communication channels embedded in biological cell membranes.

D. Organization in Statistical Mechanics

The spatial organization that spontaneously emerges in one-dimensional and two-dimensional spin systems, though arising in mere physical systems, presents a similar challenge [36]. Figure 2, from there, shows a suite of information processes as a function of spin location in lattice configurations. Specifically, it plots the local entropy density h_μ, anticipated information ρ_μ, ephemeral information r_μ, bound information b_μ, and enigmatic information q_μ across lattice sites. Reference [36] then goes on to interpret what these quantities mean in terms of the underlying physical interactions and emergent spin patterns. We return to these information signals later on, as they play a key role.

E. Approach

Looking across these very different complex adaptive systems reveals common features: (i) framing the overall system in terms of one or a group of agents and their interactions with an environment; (ii) the agent's role in monitoring the environment by extracting real-time signals—sequences of various kinds of informational measure; and (iii) the semantics in those measures—how an agent uses or interprets such signals. Figure 3 illustrates the basic measurement channel picture of the environment and agent—a picture that frames our overall development.

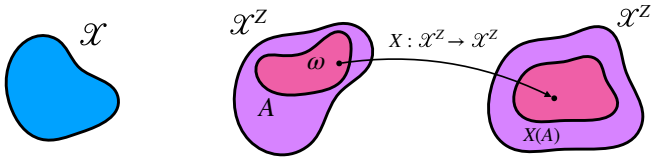

FIG. 4. Discrete-time, discrete-value stochastic process as a sequence $\mathcal{X}^{\tau}$ of indexed random variables $\{X_t : t \in \mathbb{Z}\}$ on the probability space $\left(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu\right)$ over event space $\mathcal{X}$ with realizations $\omega \in \mathcal{X}^{\mathbb{Z}}$.

Historically, calculating these informational signals has been quite theoretically and computationally demanding. Very recently, though, Ref. [24] showed how to perform these calculations using practical and efficient algorithms, progress that stimulates much of the following. Moreover, the companion Ref. [40] introduces out a thorough-going information-theoretic framing for deploying the information measures we develop in the following.

With this background and these motivations laid out, we are now ready to turn to our theoretical development.

III. STOCHASTIC PROCESSES

The following briefly summarizes discrete stochastic processes, adapting the more general framework of Ref. [41] to this simpler setting. (For measure-theoretic background see Refs. [42, 43].)

A (discrete-time, discrete-valued) *stochastic process* X consists of a sequence of indexed random variables (RVs) $\{X_t\}_{t\in\mathbb{Z}}$ defined on a probability space $\left(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu\right)$. (See Fig. 4.) X_t is the random variable representing a value $x \in \mathcal{X}$ observed at time t. $\mathcal{X}$ is the RVs' common state space or event space—the set of possible observed values or X's *alphabet*. We take it to be a finite set $\mathcal{X} = \{1, \ldots, k\}$.

A process *realization* ω is a single outcome of the stochastic process. That is, realizations are taken from the *sample space*: $\omega \in \mathcal{X}^{\mathbb{Z}}$. This is the collection of a process' behaviors. Σ is the sigma algebra that describes the process' events—possible subsets of realizations.

The measure μ is defined on measurable subsets A of realizations:

$$\mu(A) = \int_{\omega\in A} d\mu(w) \ . \tag{1}$$

for $A \in \Sigma$; that is, measurable subsets of $X^{\mathbb{Z}}$. As we detail shortly, the measure over bi-infinite realizations determines the probability of sets of realizations and of finite sequences via integration over subsets, as in Eq. (1).

The stochastic process X is generated by *cylinder sets*:

$$U_{t,w} = \{X : x_{t+1}, \ldots, x_{t+\ell} = w\} \ ,$$

where $w \in \mathcal{X}^{\ell}$ is a *word* of length ℓ. For a stationary process, the *word probabilities*:

$$\Pr(x_1 \ldots x_\ell) = \mu(U_{0,x_1\ldots x_\ell}) \tag{2}$$

are sufficient to uniquely define the measure μ.

That is, indexing is temporal and denoted by the use of subscripts $\mathbb{Z}$. It is convenient to write a realization of $\{X_t\}_{t\in\mathbb{Z}}$ as an indexed sequence. For example, we write $X_t = x$ to say that $x \in \mathcal{X}$ is the specific value of X at time t. On occasion, we shorten this to x_t. That is, uppercase (e.g., X_t) indicates the variable and lowercase (e.g., x_t) a realized value.

Blocks of consecutive RVs, called *words*, are denoted by $X_{n:m} = \{X_t : n < t \leq m; n, m \in \mathbb{Z}\}$ with the left index inclusive and the right exclusive. For example, $X_{0:3} = X_0X_1X_2$. A word may also refer to a particular realization of a given length, denoted in lowercase. For instance, one might write $x_{0:3} = x_0x_1x_2$ or $x_{0:3} = bac$, if (say) $x_t \in \mathcal{X} = \{a, b, c\}$.

A. Process Dynamics

Stochastic processes themselves can evolve over time. When working with these dynamics, we need to refer to a process relative to a particular time t. To do this we use superscripts. Denote the process referenced to time t as $X^t = \{X_t\}_{t\in\mathbb{Z}}$ and that referenced to time $t+1$ as $X^{t+1} = \{X_{t+1}\}_{t\in\mathbb{Z}}$.

A natural dynamic for a stochastic process is given by the *shift operator*: $\tau : \mathcal{X}^{\mathbb{Z}} \to \mathcal{X}^{\mathbb{Z}}$ that simply advances time $t \to t+1$:

$$(\tau X)_t = X_{t+1} \ .$$

This describes the shift acting on a process' individual RVs. For the shift acting on the entire process we have:

$$\{(\tau X)_t\}_{t\in\mathbb{X}} = \{X_{t+1}\}_{t\in\mathbb{Z}}$$

or, more simply:

$$\tau X^t = X^{t+1} \ .$$

The shift also acts on measures over $\mathcal{X}^{\mathbb{Z}}$:

$$(\tau\mu)A = \mu(\tau^{-1}A) \ ,$$

for measurable $A \in \Sigma$.

A stochastic process paired with the shift operator becomes a dynamical system $(\mathcal{X}^{\mathbb{Z}}, \Sigma, \mu, \tau)$. A stochastic process is stationary if the measure is time-shift invariant: $\tau\mu = \mu$. It is *ergodic* if, for all shift-invariant sets $\tau\mathcal{I} = \mathcal{I}$, $\mathcal{I} \subset \mathcal{X}^{\mathbb{Z}}$, either $\mu(\mathcal{I}) = 0$ or 1. Said more simply, an ergodic stochastic process cannot be decomposed into other ergodic components. (Later on, we return to notions of stationarity and ergodicity appropriate for information processes and cognitive agents.)

When considering random variables and their probabilities, we continue, as above, to denote random variables by capital Latin letters and specific realizations by lower case. For example, $\Pr(X_t) = \{\Pr(X_t = x) : x \in \{1, \ldots, k\}\}$ and $\Pr(X = x) = \mu(\{x\} \in \mathcal{X})$.

B. Sliding Window Processes

It will be helpful to select a subset of related (time local) RVs in a process via a function $h(\cdot)$ that gives offsets for the indices of those RVs:

$$h_r(t) = \{t - r, \ldots, t - 2, t - 1, 0, t + 1, t + 2, \ldots t + r\} \ .$$

In this way, one extracts a new *sliding window* stochastic process $(Y_t)_{t\in\mathbb{Z}}$ composed of a series of groups of the original process' RVs. For example, if $r = 1$:

$$\begin{aligned} Y_t &= X_{[h_1(t)]} \\ &= X_{t-1}X_tX_{t+1} \text{ and} \\ Y_{t+1} &= X_{[h_1(t+1)]} \\ &= X_tX_{t+1}X_{t+2} \ . \end{aligned}$$

C. An Experiment and Its Realizations

An agent interacts with its environment by performing an *experiment* $\mathcal{E} = \{\omega^i \in \Sigma : i = 1, 2, \ldots, M\}$ consisting of a number M of realizations ω^i. The agent prepares the environment appropriately and initiates an experimental *run* ω^i to generate an arbitrarily long realization $x_{0:\infty} = x_0, \ldots, x_\infty$, where each is a sequence of individual observations $x_t \in \mathcal{X}$. In this way, an experiment is the set $\mathcal{E} = \{\omega^1 = x'_{0:\infty}, \omega^2 = x''_{0:\infty}, \ldots, \omega^M = x'''_{0:\infty}\}$ consisting of all of the data—process realizations—available to an agent.

D. Stationary

Intuitively, the statistics of a stationary process do not depend on time.

Definition 1. In a *stationary* process the measure is time-shift invariant—$\tau\mu = \mu$—such that:

$$\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell}) \ ,$$

for all $t \in \mathbb{Z}$, $\ell \in \mathbb{Z}^+$.

For a stationary process, the index denoting the present can nominally be set to any value without altering subsequent analysis.

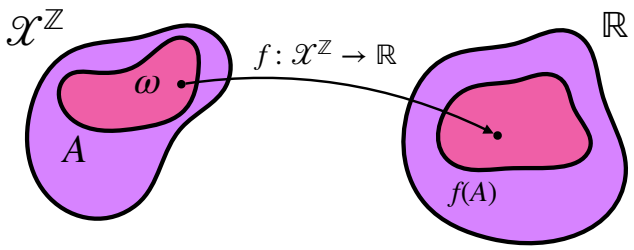

FIG. 5. Real-valued measurable function of a stochastic process: $f : \mathcal{X}^{\mathbb{Z}} \to \mathbb{R}$.

E. Ergodic

Intuitively, any realization $\omega^i \in \Sigma$ of an ergodic process has the same statistical properties as any other realization $\omega^{j\neq i}$. More formally, define the *ensemble average* $E[f]$ of a measurable function $f : X^{\mathbb{Z}} \to \mathbb{R}$:

$$E[f] = \int_{\omega\in X^{\mathbb{Z}}} f(\omega)d\mu(\omega) \ .$$

A function of a sample is often referred to as a *statistic*. And so, $E[f]$ is a statistic of the process samples ω^i for some function $f(\cdot)$.

Compare this to the time average of the function:

$$\langle f(\omega)\rangle_t = \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} f(\tau^{k-1}\omega) \ ,$$

for μ-almost every $\omega \in \mathcal{X}^{\mathbb{Z}}$.

Definition 2. An *ergodic process* consists of realizations for which time-average statistics equal ensemble averages:

$$\langle f(\omega)\rangle_t = E[f(\omega)] \ ,$$

where $f(\cdot)$ is any measurable function and for μ-almost every $\omega \in \mathcal{X}^{\mathbb{Z}}$.

F. Functions of a Stochastic Process

Recall that if Z is a random variable and $f(\cdot)$ a real-valued, measurable function, $f(Z)$ is a random variable with a real-valued event space. Thus, since probability is such a function of a random variable, $Y = \Pr(Z)$ is also a real-valued random variable with event space $\mathcal{Y} = [0, 1]$ [44]. The following relies on this observation repeatedly.

Helpfully, this observation extends to sequences $\ldots X_{-1}X_0X_1 \ldots$ of random variables that define a stochastic process $X = \{X_t : t \in \mathbb{Z}\}$. The main lesson is that a function of a stochastic process is a stochastic process. See App. A which considers real-valued functions $f(w)$ of realizations: $f : X^{\mathbb{Z}} \to \mathbb{R}$.

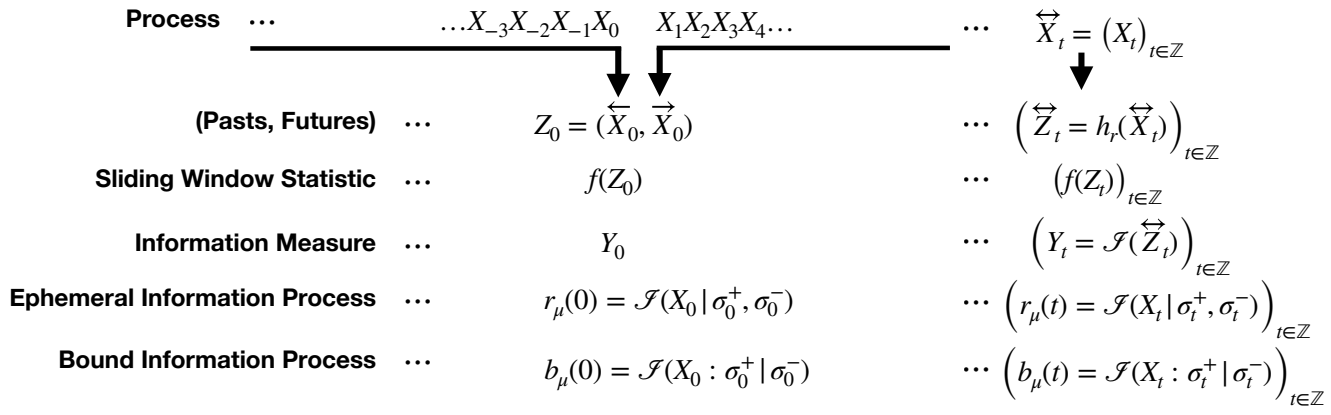

FIG. 6. Sliding window processes: The given stochastic process appears on the top line with the construction of sliding window processes from it illustrated going line by line. The last lines illustrate the ephemeral information process $r_\mu(t)$ and the bound information process $b_\mu(t)$. Recall that the forward and reverse causal states are functions of the semi-infinite past and futures, respectively—$\sigma^+ = \epsilon(\overleftarrow{X})$ and $\sigma^- = \epsilon(\overrightarrow{X})$—which aggregate RVs are available in the Z_t process.

The (push-forward) measure of a function of set $A \subset \mathcal{X}^{\mathbb{Z}}$ is:

$$\mu(f(A)) = \mu(f^{-1}(A)) \ .$$

(See Fig. 5.)

We are interested in an agent that observes a process through a finite-duration window and manipulates that information at each time. Define a *finite range* r, real-valued function $f : \mathbb{R}^{2r+1} \to \mathbb{R}$, $r < \infty$.

Proposition 1. *Applying f to a stochastic process X gives a new stochastic process $Y = \{Y_t : t \in \mathbb{Z}\}$, the components of which are:*

$$Y_t = f\left(X_{[h_r(t)]}\right) \ .$$

Proof. $Y = f(X)$'s probability space $(\mathcal{Y}, \Sigma_Y, \mu_Y) = (\mathbb{R}^{\mathbb{Z}}, \Sigma_Y, \mu_Y)$. Y's event space Σ_Y is the sigma algebra over $\mathbb{R}^{\mathbb{Z}}$. Y's measure μ_Y is given in terms of μ_X in two steps: giving the process Y_t via a function of RVs in a window onto X and then determining the measure μ_Y in terms of μ_X.

First, consider the sliding window process $Z = h_r(X)$, where h_r is a finite-range r, index function as above. That is, Z_t's event space consists of the real-valued vectors $z = (x_0, x_1, \ldots, x_{2r+1}) \in \mathbb{R}^{2r+1}$. Then the process Z is a time series of vectors $Z_t = (X_{t-r}, \ldots, X_t, \ldots X_{t+r})$.

And, second, we get Z's measure $\mu_{t,z}$ from X's μ_X:

$$\mu_{t,z} = \int_{\omega \in U_{t,z}} d\mu_X(\omega) \ ,$$

with the cylinder $U_{t,z} = \{X : (x_t, x_{t+1}, \ldots, x_{2r+1}) = z\}$ associated with vector $z = (z_0, z_1, \ldots, z_{2r+1})$. Since X is stationary we can set $t = 0$ and work with U_0, z; simplifying notation to μ_Z and U_z.

Next, we obtain the real-valued process Y_t from the vector-valued process Z_t. However, this is simply a component-wise transformation of the vector series by the real-valued function $f(\cdot)$: $Y_t = f(Z_t)$.

Finally, we determine the measure μ_Y for the real-valued process Y_t from the measure μ_Z for the vector-valued process Z_t. Recall the real-valued function $f : \mathbb{R}^{2r+1} \to \mathbb{R}$ that maps real vectors z to some real value: $f(z)$. To get Y's measure μ_Y in terms of μ_Z we integrate over the set of vectors $z \in f^{-1}(y)$, that give the real value y:

$$\mu_Y(y) = \int_{z \in f^{-1}(y)} d\mu_Z(z) \ , \tag{3}$$

where $y = \ldots y_{t-1} y_t y_{t+1} \ldots \in \mathbb{R}^{\mathbb{Z}}$ is a realization of Y. And so, $Y = f(Z)$'s probability space is $(\mathbb{R}^{\mathbb{Z}}, \Sigma_Y, \mu_Y)$.

Together, these determine Y's probability space in terms of X's probability space $(\mathcal{X}, \Sigma_X, \mu_X)$ establishing that $Y = f \circ h(X) = f(X)$ is a stochastic process with measure:

$$\mu_Y(y) = \int_{\omega \in U_{t,z}} \int_{z \in f^{-1}(y)} d\mu_Z(z) d\mu_X(\omega) \ .$$

□

Moreover, if X is stationary and ergodic, so is process Y. As the following two results establish.

Proposition 2. *If X is a stationary process and f is a finite-range, real-valued function, then $Y = f(X)$ is stationary.*

Proof. We must show that Y is time-shift invariant: $\Pr(Y_{t:t+\ell}) = \Pr(Y_{0:\ell})$, for all $t \in \mathbb{Z}$. We are given that X is time-shift invariant: $\Pr(X_{t:t+\ell}) = \Pr(X_{0:\ell})$, for all $t, \ell \in \mathbb{Z}$. And so, the process over its $2r+1$ blocks $X_{0:2r+1}$ is time-shift invariant. Since Y values are direct functions of these blocks, Y's sequences are time-shift invariant and so Y is stationary. □

Corollary 1. *If X is a stationary process, the* probability process $Y = \Pr(X)$ *is stationary.*

Proof. X's sequence probabilities $Y = \Pr(X)$ are RVs. X's stationarity guarantees that a time series of its sequence probabilities $\ldots Y_{t-1}Y_tY_{t+1}\ldots$ is time-shift invariant. However, sequence probabilities Y_t are obtained from a finite-range, real-valued function $\Pr(\cdot)$. And so, by the preceding proposition, the stochastic process $Y = \Pr(X)$ is stationary. □

Moreover:

Proposition 3. *If X is a stationary and ergodic process and f is a finite-range, real-valued function, then $Y = f(X)$ is stationary and ergodic.*

Proof. (See Ref. [42, Thm. 36.4].) We have process $Y = \{(Y)_t : t \in \mathbb{Z}\}$, with $Y_t = f\left(X_{t-r}, \ldots, X_t, \ldots, X_{t+r}\right)$. Y is stationary from the preceding proposition. We must show that Y is ergodic:

$$\langle g(Y)\rangle_t = E[g(\omega)] \ ,$$

for $\omega \in \Sigma_Y$ and functions g. We are given that this holds for the original process X:

$$\langle h(X)\rangle_t = E[h(x)] \ ,$$

for $x \in X^{\mathbb{Z}}$ and all functions h. Thus, it also holds for $h = f$:

$$\begin{aligned}\langle f(X)\rangle_t &= E[f(x)] \\ \langle Y\rangle_t &= E[\omega] \ ,\end{aligned}$$

and the expectation is taken over all $\omega \in X^{\mathbb{Z}}$ and $\Sigma_Y = f(\Sigma_X)$. □

Corollary 2. *If X is stationary and ergodic, then the probability process* $\Pr(X)$ *is stationary and ergodic.*

Proof. We are given that X's sequence probabilities are time-shift invariant and that it has only a single ergodic invariant set. Y's sequence probabilities are obtained from X via a finite-range, real-valued function $\Pr(\cdot)$. And so, the preceding proposition shows that the process $\Pr(X)$ is stationary and ergodic. □

Figure 6 illustrates the construction of the sliding window processes from which information processes are constructed.

IV. INFORMATIONS AND INFORMATION PROCESSES

We introduce Shannon's information measures, information diagrams, and information processes, showing how to use them to describe a variety of informational properties of a stochastic process X. Table I provides a roadmap for the various kinds of stochastic informational processes the following introduces. The companion, Table II, lists several information processes that are time symmetric in the sense that the sets of RVs from which they are formed is symmetric under $t \to -t$.

⇐ Reverse		Processes			Forward ⇒
Reverse Statistical	$\cdots$	$\mathfrak{I}\left[\sigma^-_{-1}\right]$	$\mathfrak{I}\left[\sigma^-_{0}\right]$	$\mathfrak{I}\left[\sigma^-_{1}\right]$	$\cdots$
Complexity	$\cdots$	$C^-_\mu(-1)$	$C^-_\mu(0)$	$C^-_\mu(1)$	$\cdots$
Reverse	$\cdots$	$\epsilon^-(\overrightarrow{X}_{-1})$	$\epsilon^-(\overrightarrow{X}_{0})$	$\epsilon^-(\overrightarrow{X}_{1})$	$\cdots$
Causal Process	$\cdots$	σ^-_{-1}	σ^-_{0}	σ^-_{1}	$\cdots$
Futures	$\cdots$	$\overrightarrow{X}_{-1}$	$\overrightarrow{X}_{0}$	$\overrightarrow{X}_{1}$	$\cdots$
Presents	$\cdots$	X_{-1}	X_0	X_1	$\cdots$
Pasts	$\cdots$	$\overleftarrow{X}_{-1}$	$\overleftarrow{X}_{0}$	$\overleftarrow{X}_{1}$	$\cdots$
Forward	$\cdots$	$\epsilon^+(\overleftarrow{X}_{-1})$	$\epsilon^+(\overleftarrow{X}_{0})$	$\epsilon^+(\overleftarrow{X}_{1})$	$\cdots$
Causal Process	$\cdots$	σ^+_{-1}	σ^+_{0}	σ^+_{1}	$\cdots$
Forward Statistical	$\cdots$	$\mathfrak{I}\left[\sigma^+_{-1}\right]$	$\mathfrak{I}\left[\sigma^+_{0}\right]$	$\mathfrak{I}\left[\sigma^+_{1}\right]$	$\cdots$
Complexity	$\cdots$	$C^+_\mu(-1)$	$C^+_\mu(0)$	$C^+_\mu(1)$	$\cdots$

TABLE I. Information processes: The given stochastic process appears on the line with Presents. All other lines are information processes that derive from it.

A. Self-Informations and Measures

The most basic quantity in information theory is the *self-information*—the amount of information learned observing a single measurement value x of a random variable X_t [45, 46]:

$$\mathfrak{i}\left[X_t = x\right] = -\log_2 \Pr(X_t = x) \ . \tag{4}$$

On the one hand, if the occurrence of x is certain, an observer learns no information and $\mathfrak{i} = 0$. On the other hand, if all realizations are equally uncertain $\Pr(X_t = x) = 1/k$ for each x, the observer gains the maximum amount of information—that is, any one occurrence is maximally informative and $\mathfrak{i}\left[X_t = x\right] = \log_2 k$.

Several comments on notation and terminology are useful initially. The prefix "self-" connotes a quantity describing an individual event's occurrence. This deviates from the typically-rare use of "self-" in "self-information" as the mutual information of a random variable with itself, which is simply its entropy: $I[X : X] = H[X]$ in the notation of Ref. [46, p. 21]. In addition, here within the theory of information measures [25], there is only a single quantity *information*, denoted $\mathfrak{I}$. In this, one has $\mathfrak{I}\left[X : X\right] = \mathfrak{I}\left[X\right]$, highlighting the redundancy in the conventional use of distinct symbols H and I. Different kinds of information—joint, conditional, mutual—are then denoted by the form of $\mathfrak{I}$'s arguments. This leads to the practical consequence that we use only the single notation $\mathfrak{I}[\cdot]$ rather than, as in basic information theory [46], using both $H[\cdot]$ and $I[\cdot]$.

Processes					
Prediction	$\cdots$	$\mathfrak{I}\left[X_{-1} \mid \sigma_{-1}^{+}\right]$	$\mathfrak{I}\left[X_0 \mid \sigma_0^{+}\right]$	$\mathfrak{I}\left[X_1 \mid \sigma_1^{+}\right]$	$\cdots$
	$\cdots$	$h_\mu^{+}(-1)$	$h_\mu^{+}(0)$	$h_\mu^{+}(1)$	$\cdots$
Predictable	$\cdots$	$\mathfrak{I}\left[\sigma_{-1}^{+} : \sigma_{-1}^{-}\right]$	$\mathfrak{I}\left[\sigma_0^{+} : \sigma_0^{-}\right]$	$\mathfrak{I}\left[\sigma_1^{+} : \sigma_1^{-}\right]$	$\cdots$
	$\cdots$	$\boldsymbol{E}(-1)$	$\boldsymbol{E}(0)$	$\boldsymbol{E}(1)$	$\cdots$
Ephemeral	$\cdots$	$r_\mu(-1)$	$r_\mu(0)$	$r_\mu(1)$	$\cdots$
Bound	$\cdots$	$b_\mu(-1)$	$b_\mu(0)$	$b_\mu(1)$	$\cdots$

TABLE II. Time Symmetric Information Processes: Derived from information measures that are symmetric in time.

With reference to self-information, the more familiar quantity from information theory, though, is its ensemble average—the Shannon entropy $\mathfrak{I}[X] = E[\mathfrak{i}[X_t]]$ of the random variable X_t:

$$\begin{aligned}\mathfrak{I}[X_t] &= E[\mathfrak{i}[X_t]] \\ &= -\sum_{x \in \mathcal{X}} \Pr(X_t = x) \log_2 \Pr(X_t = x) \ . \end{aligned} \quad (5)$$

This requires being given or knowing ahead of time the single-symbol average probabilities $\{\Pr(X_t = x), x \in \mathcal{X}\}$. Given a stochastic process, it can also be developed from measuring the self-information over time:

$$\langle \mathfrak{i}[X_t = x] \rangle = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \mathfrak{i}[X_t = x] \ ,$$

assuming that the process is stationary and ergodic.

And, from this, the ensemble average is:

$$\mathfrak{I}[X_t] = -\sum_{x \in \mathcal{X}} \langle \mathfrak{i}[X = x] \rangle \ .$$

Similarly, there is the relationship between a pair of jointly-distributed random variables, say, X_t and $X_{t'}$. To simplify, considered time-delayed RVs where $t' = t + \tau$. The *joint entropy* $\mathfrak{I}[X_t, X_{t'}](\tau)$ is of the same functional form as Eq. (5), applied to the joint distribution $\Pr(X_t, X_{t'})$.

First, we have the joint self-information:

$$\mathfrak{i}[X_t = x, X_{t'} = x'](\tau) = -\log_2 \Pr(X_t = x, X_{t'} = x') \ , \quad (6)$$

for $x, x' \in \mathcal{X}$. Second, the ensemble-averaged version:

$$\mathfrak{I}[X, X'](\tau) = -\sum_{x, x' \in \mathcal{X}} \Pr(X_t = x, X_{t'} = x') \log_2 \Pr(X_t = x, X_{t'} = x') \ . \quad (7)$$

For which we can also implement a time-averaged quantity:

$$\langle \mathfrak{i}[X_t = x, X_{t'} = x'] \rangle(\tau) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \mathfrak{i}[X_t = x, X_{t'} = x'] \ ,$$

From this, the ensemble average is then given:

$$\mathfrak{I}[X, X'](\tau) = -\sum_{x, x' \in \mathcal{X}} \langle \mathfrak{i}[X_t = x, X_{t'} = x'] \rangle(\tau) \ .$$

Effectively, via sampling, the time average "empirically" provides an estimate of the joint $\Pr(X_t = x, X_{t'} = x')$.

These can be straightforwardly extended, in principle, from pairs of variables to the *multivariate joint entropy* $\mathfrak{I}(\mathfrak{X})$ of a set of N variables $\mathfrak{X} = \{X_i \mid i \in (1, \ldots, N)\}$. (In this, the temporal relationship between the variables needs to be specified by an appropriate set of indexes; such as, time delays.)

Finally, the *conditional entropy* $\mathfrak{I}[X_t \mid X_t'](\tau)$ is of a similar functional form as above, but applied to the conditional distribution $\Pr(X_t | X_{t'})$. It gives the information learned from observing a random variable X_t at one time given knowledge of random variable $X_{t'}$ at another $t' = t + \tau$. We can write this in terms of the joint entropy:

$$\mathfrak{I}[X_t \mid X_{t'}](\tau) = \mathfrak{I}[X_t, X_{t'}](\tau) - \mathfrak{I}[X_{t'}] \ . \quad (8)$$

First, if we break this down as above, we have the conditional self-information:

$$\mathfrak{i}[X_t = x \mid X_{t'} = x'] = -\log_2 \Pr(X_t = x | X_{t'} = x') \ , \quad (9)$$

for $x, x' \in \mathcal{X}$. Second, there is the ensemble-averaged version:

$$\mathfrak{I}[X \mid X'](\tau) = -\sum_{x, x' \in \mathcal{X}} \Pr(X_t = x, X_{t'} = x') \log_2 \Pr(X_t = x | X_{t'} = x') \ . \quad (10)$$

For which we can also implement a time-averaged quantity:

$$\begin{aligned}&\langle \mathfrak{i}[X_t = x \mid X_{t'} = x'] \rangle(\tau) \\ &\quad = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \mathfrak{i}[X_t = x \mid X_{t'} = x')] \ ,\end{aligned}$$

If the stochastic process is stationary and ergodic, then the ensemble average is given:

$$\mathfrak{I}[X \mid X'](\tau) = -\sum_{x, x' \in \mathcal{X}} \langle \mathfrak{i}[X_t = x \mid X_{t'} = x')] \rangle(\tau) \ .$$

The information shared between random variables is the *mutual information*. The mutual self-information of observing x and x' at times t and (delayed) t', respectively, is:

$$\mathfrak{i}[X_t = x : X_{t'} = x'] = \log_2 \left(\frac{\Pr(X_t = x, X_{t'} = x')}{\Pr(X_t = x) \Pr(X_{t'} = x')} \right) \ . \quad (11)$$

When appropriately ensemble-averaged this recovers the version familiar from elementary information theory:

$$\begin{aligned}&\mathfrak{I}\left[X_t : X_{t'}\right](\tau)\\&= \sum_{\substack{x' \in \mathcal{X} \\ x \in \mathcal{X}}} \Pr\left(X_t = x, X_{t'} = x'\right) \mathfrak{i}\left[X_t = x : X_{t'} = x'\right] . \quad (12)\end{aligned}$$

And, a time-averaged version is:

$$\begin{aligned}&\langle \mathfrak{i}\left[X_t = x : X_{t'+\tau} = x'\right] \rangle(\tau)\\&\quad = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \mathfrak{i}\left[X_t = x : X_{t'} = x'\right] . \quad (13)\end{aligned}$$

Again, if the stochastic process is stationary and ergodic, then the time-average and ensemble average are equal. And, the ensemble average is given:

$$\mathfrak{I}\left[X_t : X_{t'}\right](\tau) = - \sum_{x, x' \in \mathcal{X}} \langle \mathfrak{i}\left[X_t = x : X_{t'+\tau} = x'\right] \rangle(\tau) .$$

The mutual information $\mathrm{I}\left[X_{m,n} : X_{p,q} : X_{r,s}\right]$ between three random-variable blocks—known as the *interaction information* or as one of the *multivariate mutual informations*—is given by the difference between mutual information and conditional mutual information:

$$\begin{aligned}&\mathrm{I}\left[X_{m,n} : X_{p,q} : X_{r,s}\right]\\&\quad = \mathrm{I}\left[X_{m,n} : X_{r,s}\right] - \mathrm{I}\left[X_{m,n} : X_{p,q} \mid X_{r,s}\right] . \quad (14)\end{aligned}$$

Tracking three-way interactions in this way between variables X, Y, and Z, say, brings up two interpretations worth highlighting here. Two variables X and Y can have positive mutual information but be conditionally independent in the presence of Z, in which case the interaction information is positive. It is also possible, though, for two independent variables to become correlated in the presence of Z, making the conditional mutual information positive and the interaction information negative.

In other words, conditioning on a third variable Z can either increase or decrease mutual information and the X and Y variables can appear more or less dependent given additional observations [46]. That is, there can be *conditional independence* or *conditional dependence* between a pair of random variables. Note that the interaction information is symmetric, so these interpretations hold regardless of the conditioning variable selected.

Finally, note that the self-information functions $\mathfrak{i}(\cdot)$ above:

1. $-\log_2 \Pr(X_t)$,
2. $-\log_2 \Pr(X_t, X_{t'})$,
3. $-\log_2 \Pr(X_t | X_{t'})$, and
4. $-\log_2 \left(\Pr(X_t X_{t'}) / \Pr(X_t) \Pr(X_{t'})\right)$.

are used in weighted averages to give ensemble averages. Their temporal and ensemble averages are equal when the process in question is stationary and ergodic.

In this, these self-informations lead to distinct and complementary informational quantities useful in characterizing a process' various properties. (This will become apparent shortly.) And, in the role they play in that averaging, they are occasionally (suggestively) referred to as "densities" [47]. This language parallels that found in continuous random variable theory whose "densities" are integrated to obtain probability "distributions".

Note that the second pair of self-informations are linear functions of the first two:

$$-\log_2 \Pr(X_t | X_{t'}) = -\log_2 \Pr(X_t, X_{t'}) - \log_2 \Pr(X_t)$$

and:

$$\begin{aligned}&-\log_2 \left(\Pr(X_t X_{t'}) / \Pr(X_t) \Pr(X_{t'})\right)\\&\quad = -\log_2 \Pr(X_t, X_{t'}) + \log_2 \Pr(X_t) + \log_2 \Pr(X_{t'}) .\end{aligned}$$

This also holds true [25, Sec. 3] for other Shannon measures that we develop shortly. And so, in the following we need only consider the ergodic properties of the marginal and joint temporal self-informations, due to the linearity of the conditions for stationarity and ergodicity.

Taken altogether, the preceding simply reviewed quantities from elementary information theory, while pointing out how they are developed from time-dependent self-informations (the "densities") via ensemble or temporal averaging. These densities play a key role later when describing how an agent, interacting in real-time with a complex environment, generates a time series—a stochastic process of—time-local self-informations.

B. Temporal Information Atoms

One of the most basic operations of a stochastic process is to create information [48, 49]—measured by the entropy density (or rate) h_μ. Importantly, in most settings there is more going on than creation. Reference [50] showed that created information consists of two parts: ephemeral information r_μ is that part of the created information which the present throws away and the bound information b_μ is that part which the process remembers—that can affect the future [51]. Said simply, processes *both* forget and remember portions of the information created at each time: $h_\mu = r_\mu + b_\mu$. This realization led to a more detailed decomposition and identification of the kinds of information and transformations with which stochastic processes operate.

Beyond time series, in the more general circumstance of a set of RVs, information can be generally decomposed into a collection of atoms; see Ref. [52]. In the time series setting, though, the RVs are related by a time shift and this simplifies understanding the various kinds of information available in a stochastic process. We refer to the observation $X_t = x$ at time t as occurring in the *present*. We call the semi-infinite sequence $X_{-\infty:t}$ the *past* at time t, which we also (more frequently) denote with a left pointing arrow $\overleftarrow{X}_t$. Accordingly, the semi-infinite

sequence $X_{t+1:\infty}$ is called the *future* at time t and denoted $\overrightarrow{X}_t$. Thus, our strategy for analyzing the information theory of stochastic processes is primarily concerned with delineating the relationships between the past, present, and future. In this, we follow Ref. [24].

Given this strategy, a useful perspective on a stochastic process is to picture it as a communication channel transmitting information from the past $\overleftarrow{X}_{t-1} = \ldots X_{t-3}, X_{t-2}, X_{t-1}$ to the future $\overrightarrow{X}_t = X_{t+1}, X_{t+2}, X_{t+3} \ldots$ through the medium of the present X_t. The past and the future are depicted in Fig. 7's *information diagram* as extending to the left and the right, respectively, to emphasize visualizing the bi-infinite chain of random variables [24].

One might expect increasing difficulty when moving from one or several isolated random variables to a stochastic process of semi-infinite random-variable chains. However and, at least initially, profiling a process' information atoms in terms of its past $\overleftarrow{X}_t$, present X_t, and future $\overrightarrow{X}_t$ requires little more information-theoretic set-up than that already given; see also Ref. [24].

Given these three random variables in play at each time t, there is a set of eight quantities—*information atoms*. These are shown in information-diagram form in Fig. 7(left) and are collected in Table III. The related aggregate measures, including h_μ already described, are also listed. Each is interpreted in more detail shortly.

However, we glossed over a rather important point—that the information measures listed above are nominally functions over joint distributions of infinite, composite variables; specifically, semi-infinite pasts and futures come into play. And, this presents several technical issues to address.

First, note that the entropy of a sequence of infinitely many random variables comprising a stationary stochastic process will either diverge or be finite. To see this, consider the block entropy $\mathfrak{I}[X_{0:\ell}]$ as a function of ℓ. It diverges when the process has positive conditional entropy $\mathfrak{I}\left[X_0 \mid \overleftarrow{X}_1\right]$. In contrast, when the conditional entropy is zero for each measurement, which occurs for periodic signals, the past and the future are exactly predictable from any measurement and so $\mathfrak{I}[X_{0:\infty}]$ contains finite information. These observations apply to when working with informational quantities over RV blocks.

(Notational reminder: The kind of information—joint, conditional, or mutual—is conveyed by the arguments to $\mathfrak{I}[\cdot]$. Thus, the quantity first listed in the previous paragraph is a joint entropy over the length-ℓ RV block $X_{0:\ell}$; whereas the second is a conditional entropy $\mathfrak{I}[X \mid Y]$. When the vertical bar there is replaced with a colon one has a mutual information $\mathfrak{I}[X : Y]$. In addition, when we are only interested in block length, we do not index off of time t, but set $t = 0$ which is allowed by stationarity, and then use ℓ to denote block length for a block located at any time t.)

Second, we assume that all other information atoms are finite. To show this, define the *excess entropy* or the total

Information Atoms		
A	$\overline{A}$	$\mathfrak{I}[A\|\overline{A}]$
$\left\{X_t\right\}$	$\left\{\overleftarrow{X}_t, \overrightarrow{X}_t\right\}$	$r_\mu(t)$
$\left\{\overleftarrow{X}_t : X_t\right\}$	$\left\{\overrightarrow{X}_t\right\}$	$b_\mu^-(t)$
$\left\{\overleftarrow{X}_t : \overrightarrow{X}_t\right\}$	$\left\{X_t\right\}$	$\sigma_\mu(t)$
$\left\{X_t : \overrightarrow{X}_t\right\}$	$\left\{\overleftarrow{X}_t\right\}$	$b_\mu^+(t)$
$\left\{\overleftarrow{X}_t : X_t : \overrightarrow{X}_t\right\}$		$q_\mu(t)$

Composite Information Measures		
A	$\overline{A}$	$\mathfrak{I}[A\|\overline{A}]$ or $\mathfrak{I}[A : \overline{A}]$ (*)
$\{X_t\}$	$\left\{\overleftarrow{X}_t\right\}$	$h_\mu^+(t) = r_\mu(t) + b_\mu^+(t)$
$\{X_t\}$	$\left\{\overrightarrow{X}_t\right\}$	$h_\mu^-(t) = r_\mu(t) + b_\mu^-(t)$
$\{X_t\}$	$\left\{\overleftarrow{X}_t\right\}$	$\rho_\mu^+(t) = q_\mu(t) + b_\mu^-(t)$ (*)
$\{X_t\}$	$\left\{\overrightarrow{X}_t\right\}$	$\rho_\mu^-(t) = q_\mu(t) + b_\mu^+(t)$ (*)
$\left\{\overleftarrow{X}_t : \overrightarrow{X}_t\right\}$		$\boldsymbol{E}(t) = b_\mu^+(t) + \sigma_\mu(t) + q_\mu(t)$ (*)
		$\boldsymbol{E}(t) = b_\mu^-(t) + \sigma_\mu(t) + q_\mu(t)$ (*)

TABLE III. (Top) Generator set (temporal information atoms) $\mathfrak{I}[A|\overline{A}]$ for the set of past, present, and future random variables $\mathfrak{X} = \left\{\overleftarrow{X}_t, X_t, \overrightarrow{X}_t\right\}$. Compare the list of $\mathfrak{I}[A|\overline{A}]$ to the areas of the information diagram depicted in Fig. 7(left). (After Ref. [24].) Note that the forward and reverse entropy rates—$h_\mu^+(t)$ and $h_\mu^+(t)$, respectively—are equal for stationary processes. (Bottom) Composite information measures h_μ^+, h_μ^-, and $\boldsymbol{E}$; see Refs. [50, 51, 53].

amount of information shared between the past and the future of the process:

$$\boldsymbol{E} = \mathfrak{I}\left[\overleftarrow{X} : \overrightarrow{X}\right] \ , \tag{15}$$

where in this case the present X_0 is taken as the start of the future $\overrightarrow{X}$. Note that this quantity is equivalent to the sum of the last three information atoms; see the last two rows in the Table III.

There is a corresponding instantaneous excess entropy at each moment:

$$\boldsymbol{E}(t) = \mathfrak{I}\left[\overleftarrow{X}_t : \overrightarrow{X}_t\right] \ , \tag{16}$$

Note that for stationary, ergodic processes $\boldsymbol{E} = \langle \boldsymbol{E}(t) \rangle_t$—the temporal average.

And, it has a corresponding self-information:

$$\boldsymbol{E}(\overleftrightarrow{X}_t) = -\log_2 \left(\frac{\Pr(\overleftarrow{X}_t, \overrightarrow{X}_t)}{\Pr(\overleftarrow{X}_t)\Pr(\overrightarrow{X}_t)} \right) \ ,$$

Again, this self-information when setting block delay $\tau = 0$ and time-averaged inside Eq. (12) becomes the familiar averaged self-excess entropy—the average past-future mutual information Eq. (15).

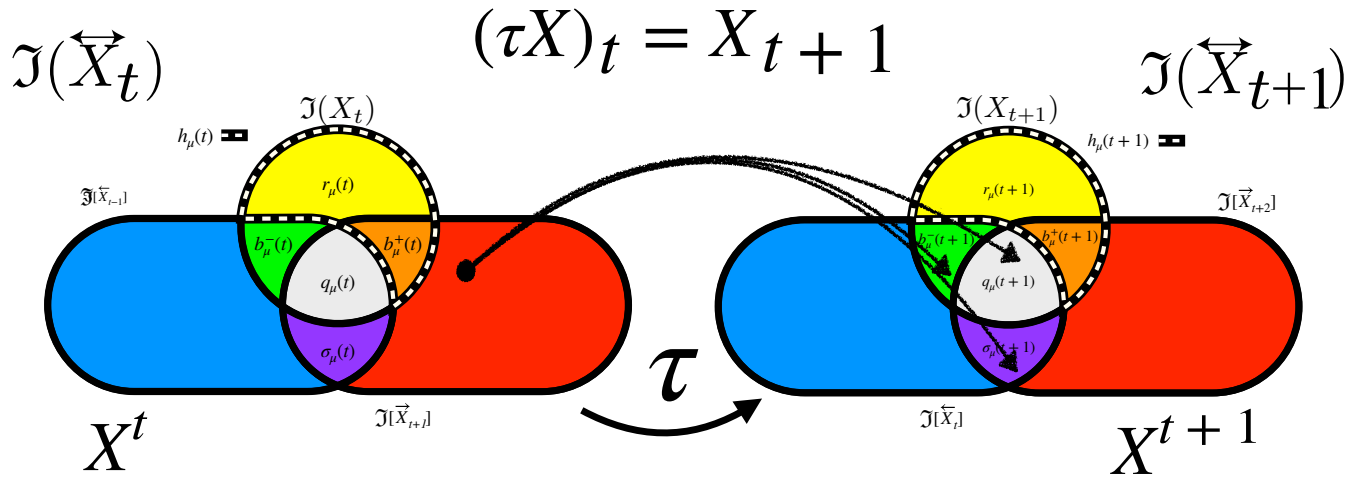

FIG. 7. Information processes $\mathfrak{I}[A|\overline{A}]$ when shift operator τ (Eq. (3)) acts on stationary, ergodic stochastic process $\{X_t\}_{t\in\mathbb{Z}}$ for atoms A in Table III. Temporal information diagrams representing how the informational relationships between the stochastic process indexed at t evolve into those indexed to $t+1$: $X^{t+1} = \tau(X^t)$—how the future $\overrightarrow{X}_t$, the present X_t, and the past $\overleftarrow{X}_{t-1}$ evolve over time to the next future $\overrightarrow{X}_{t+2}$, present X_{t+1}, and past $\overleftarrow{X}_t$. (Left) I-diagram at time t, formally denoted $\mathfrak{I}(\overleftrightarrow{X}_t)$; (Right) I-diagram at time $t+1$, formally denoted $\mathfrak{I}(\overleftrightarrow{X}_{t+1})$. The i-diagrams are labeled with the time t and $t+1$ ephemeral informations $r_\mu(t)$ and $r_\mu(t+1)$, the binding informations $b_\mu^\pm(t)$ and $b_\mu^\pm(t+1)$, the enigmatic informations $q_\mu(t)$ and $q_\mu(t+1)$, the elusive informations $\sigma_\mu(t)$ and $\sigma_\mu(t+1)$. The Shannon entropy rates $h_\mu(t) = r_\mu(t) + b_\mu(t)$ and $h_\mu(t+1) = r_\mu(t+1) + b_\mu(t+1)$—composite information atoms—are outlined with a white-dashed line.

Processes with finite excess entropy $\boldsymbol{E}$ (Eq. (15)) are called *finitary processes*. We will assume here that all processes under consideration are finitary, and therefore all other information atoms of the process are also finite, except for the two over semi-infinite pasts and futures. We can see this by noting the information in the present X_t must be finite, as it is a single measurement and is bounded by $\log_2|\mathcal{X}|$. So, any information quantity containing X_t must be finite—all atoms, except semi-infinite groups, must be less than or equal to the excess entropy.

Third, each of the finite atomic information quantities may be expressed as the asymptotic growth rate of a block information quantity [50]. That is to say, the growth rate of an information quantity taken over words of length ℓ as ℓ goes to infinity.

For example, consider the total amount of information contained in words of length ℓ. As ℓ increases, this block entropy scales as:

$$\mathfrak{I}[X_{0:\ell}] \sim \boldsymbol{E} + \ell h_\mu, \text{ as } \ell \to \infty,$$

where h_μ is the *entropy rate* and is defined:

$$h_\mu = \lim_{\ell\to\infty} \frac{\mathfrak{I}[X_{0:\ell}]}{\ell} \ . \tag{17}$$

In some applications, this form is also referred to as a *density*.

It can be shown that this asymptotic quantity is equivalent to the entropy in the present conditioned on the past:

$$\begin{aligned} h_\mu &= \lim_{\ell\to\infty} \mathfrak{I}[X_0 \mid X_{-\ell:0}] \\ &= \mathfrak{I}\left[X_0 \mid \overleftarrow{X}\right] \ . \end{aligned} \tag{18}$$

This expression gives an operational view of the information measure—the entropy rate is the amount of new information learned upon a single new observation of the process. This also allows us to identify the quantity on the process i-diagram in Fig. 7(left).

There is also the time-local version:

$$h_\mu(t) = \mathfrak{I}\left[X_t \mid \overleftarrow{X}_t\right]$$

and a self-information version:

$$h_\mu(t,x) = \mathfrak{i}\left[X_t = x \mid \overleftarrow{X}_t\right] \ .$$

The *anticipated information* ρ_μ is the complement of the new information h_μ—future information that can be predicted from the past:

$$\rho_\mu = \mathfrak{I}[X_0] - h_\mu \tag{19}$$

$$= \mathfrak{I}\left[X_0 : \overleftarrow{X}\right] \ . \tag{20}$$

And, there is the time-local version:

$$\rho_\mu(t) = \mathfrak{I}\left[X_t : \overleftarrow{X}_t\right]$$

and a self-information version:

$$\rho_\mu(t,x) = \mathbf{i}\left[X_t = x : \overleftarrow{X}_t\right] \ .$$

Depending on the semi-infinite past or future, the entropy rate and excess entropy are global measures of the information contained in a process. Specifically, h_μ is the rate of information created and $\boldsymbol{E}$ is that portion of created information communicated from the past to the future. The entropy rate is useful because of its immediate interpretability. And, also as it can be shown to be equivalent to the Kolmogorov-Sinai entropy, a dynamical invariant of the underlying system (environment).

C. General Atoms

We can identify other atomic self-information densities for the information atoms given in Table III. These have proved themselves to be useful. We will describe these quantities and give their self-information versions and their intuitive meanings here. They play a key role in interpreting information processing in Sec. VII's example processes. The interested reader should consult Ref. [50] for a deeper discussion.

Ephemeral information rate r_μ That information localized to an isolated variable, but not correlated to its peers. In other words, the rate can be interpreted as the average information in a single measurement of a process undetermined by any temporal correlation:

$$r_\mu(t) = \mathfrak{I}\left[X_t \mid \overleftarrow{X}_t, \overrightarrow{X}_t\right] \ .$$

The self-information version is:

$$r_\mu(t,x) = \mathbf{i}\left[X_t = x \mid \overleftarrow{X}_t, \overrightarrow{X}_t\right] \ .$$

Binding rate b_μ The rate of increase of the total information in a block minus the ephemeral entropy. It is also the rate at which a process stores created information. There are two equivalent quantities, forward binding rate b_μ^+ and reverse binding rate b_μ^-. For stationary processes we always have $b_\mu^+ = b_\mu^-$. The forward and reverse binding rates can be interpreted as how correlated any given measurement of a process is with the future and the past, respectively:

$$b_\mu^+(t) = \mathfrak{I}\left[X_t, \overrightarrow{X}_t \mid \overleftarrow{X}_t\right] \text{ and} \tag{21}$$

$$b_\mu^-(t) = \mathfrak{I}\left[X_t, \overleftarrow{X}_t \mid \overrightarrow{X}_t\right] \ . \tag{22}$$

The two self-information versions are:

$$b_\mu^+(t,x) = \mathbf{i}\left[X_t = x, \overrightarrow{X}_t \mid \overleftarrow{X}_t\right] \text{ and}$$

$$b_\mu^-(t,x) = \mathbf{i}\left[X_t = x, \overleftarrow{X}_t \mid \overrightarrow{X}_t\right] \ .$$

Enigmatic rate q_μ The interaction information between any given measurement of a process and the infinite past and future:

$$q_\mu(t) = \mathfrak{I}\left[X_t; \overleftarrow{X}_t; \overrightarrow{X}_t\right] \ . \tag{23}$$

As this is a multivariate mutual information, it can be negative. The self-information version is:

$$q_\mu(t,x) = \mathbf{i}\left[X_t = x; \overleftarrow{X}_t; \overrightarrow{X}_t\right] \ .$$

Elusive information σ_μ The amount of information shared between the past and future that is not communicated through the present. The elusive information is also sometimes called *state information*, as it represents the amount of information the observer needs to access beyond measurement to build a predictive model of the system:

$$\sigma_\mu(t) = \mathfrak{I}\left[\overleftarrow{X}_t : \overrightarrow{X}_t \mid X_t\right] \ . \tag{24}$$

The self-information version is:

$$\sigma_\mu(t,x) = \mathbf{i}\left[\overleftarrow{X}_t, \overrightarrow{X}_t \mid X_t = x\right] \ .$$

Interpretations There are a number of useful relationships. Let's mention just one and recommend the references above for fuller discussion.

The ensemble-averaged forward binding information and ephemeral information taken together are equal to the ensemble-average entropy rate at each time t:

$$h_\mu(t) = r_\mu(t) + b_\mu^+(t) \ .$$

This decomposes the newly-created information per measurement into a part correlated with the future and a part only correlated with the current measurement and "forgotten" in the next time step. To take one example, $b_\mu(t)$ is the instantaneous rate of storing information internally for use at some future time.

The ensemble-averaged reverse binding information and enigmatic information taken together equal the ensemble-average forward anticipation rate at each time t:

$$\rho_\mu^+(t) = q_\mu(t) + b_\mu^-(t) \ .$$

This decomposes the predictable information per measurement into a part correlated with the past and a part only correlated with the current measurement. Similarly for:

$$\rho_\mu^-(t) = q_\mu(t) + b_\mu^+(t) \ .$$

These informations are depicted in visual form by the i-diagram in Fig. 7(left) and explicitly (minus the "endcap" atoms) in Table III. In principle, it is possible to profile any process (or environment) using these atoms to give a full picture of its informational structure.

D. Dynamic Information Processes

Implicit in detailing the time-dependence of information atoms is an agent that moment-by-moment interprets an observation of its environment as containing useless or useful information—used in predictions, decisions, or actions. Such time-dependent information measures are quite common—as Sec. II noted—although perhaps not always called out as such dynamic quantities.

There are even pitfalls in interpretations. For example, the *transfer entropy* [54]—a time-dependent measure of mutual information within a sliding window, was introduced to detect causal interactions in dynamical systems. The causation entropy [55], also time dependent, was then introduced to address several of its weakness in conflating conditional dependence and independence [56]. Despite such concerns, these causal-detection measures have come to be widely used in the empirical sciences for structural inferences. This, at least, attests to the need for such statistics.

Taken altogether, though, the foregoing lays out the main setting in which such time-dependent statistics operate. With this, it is now time to be more explicit about the agent and its functioning.

To set this up, though, we must first address the time evolution of information measures, which has only been indirectly specified thus far via the shift operator τ of Eq. (3). Figure 7 makes this explicit by contrasting the i-diagrams of measures at time t (left) and those at the next moment $t+1$ (right)—that is, measures for processes X^t and X^{t+1}, respectively.

At time t, we have measures of the past, present, and future; viz., $\mathfrak{I}\left[\overleftarrow{X}_{t-1}\right]$, $\mathfrak{I}\left[X_t\right]$, and $\mathfrak{I}\left[\overrightarrow{X}_{t+1}\right]$, accompanied by the atoms $b_\mu^-(t)$, $r_\mu(t)$, $b_\mu^+(t)$, $q_\mu(t)$, and $\sigma_\mu(t)$. At the next time, we have $\mathfrak{I}\left[\overleftarrow{X}_{t}\right]$, $\mathfrak{I}\left[X_{t+1}\right]$, and $\mathfrak{I}\left[\overrightarrow{X}_{t+2}\right]$ and their companions $b_\mu^-(t+1)$, $r_\mu(t+1)$, $b_\mu^+(t+1)$, $q_\mu(t+1)$, and $\sigma_\mu(t+1)$.

At each time step, τ transforms the measures at t into those at time $t+1$, with the former typically splitting and being shared across the latter. One such atomic transformation is depicted: $\mathfrak{I}\left[\overrightarrow{X}_{t+1}\right]$ maps to $b_\mu^-(t+1)$, $q_\mu(t+1)$, and $\sigma_\mu(t+1)$. That is, the time-t forward block entropy splits and contributes to the time $t+1$ reverse bound information, enigmatic information, and elusive information.

Reference [40] provides fuller details and interpretations in other cases. All in all, though, this gives a view of the dynamic operation of an information process—extracting, storing, and processing various kinds of information. Section VII gives an even more explicit view of dynamical information processing in when analyzing the information process time series generated by several examples.

As we are working in the setting of temporal or pointwise information measures, a final cautionary remark is in order. While motivated by elementary information theory for a given set of random variables, there are important interpretational differences for information processes.

Concretely, the elementary information theory of two random variables leads to informations that are positive—entropy, entropy rate, information gain, mutual information, and so on. It is well-known, however, that when applied to three random variables the mutual information can become negative. This corresponds to the previously-mentioned phenomena of conditional dependence versus conditional independence induced between two random variables by a third [46, Prob. 2.25]. The key point here is that a similar deviation from intuitions derived from elementary information theory occurs for pointwise or temporal information measures.

To be specific consider the information processes consisting of residual information $\rho_\mu(t)$ and bound information $b_\mu(t)$. As conditional mutual informations they can be negative in the pointwise setting. For $\rho_\mu(t)$, which is a simple mutual information, negativity appears when its associated self-information argument $\Pr\left(\overleftarrow{X}(t), X_t\right) / \Pr\left(\overleftarrow{X}(t)\right) \Pr\left(X_t\right) < 1$ (or when $\Pr\left(\overleftarrow{X}(t), X_t\right) < \Pr\left(\overleftarrow{X}(t)\right) \Pr\left(X_t\right)$). This occurs in the pointwise setting when the two observations—X_t and $\overleftarrow{X}(t)$—occur less often than one would expect if the process were independent, identically distributed. Whereas, $\rho_\mu(t) > 0$ implies the past and present occur more frequently than one expects if they were independent. Section VII provides numerous examples of this and the negativity of other temporal information measures that one would, without warning, assume positive. The practical result is a need to retool intuitions from elementary information theory when interpreting temporal informations.

V. COGNITIVE AGENTS

The following introduces a more explicit notion of (i) an agent interacting with (ii) an environment whose behavior the agent monitors by making sequential observations, interpreting the latter to, perhaps, make decisions in the service of future interactions.

We refer to the sequential measurement and interpretation as *cognition* and the observers as *cognitive agents*. Granted this invokes a rather literal notion of cognition—one that falls short of that in animals say. That said, the goal in the following is to lay out the basic dynamical and informational foundations for cognitive agents of any stripe. The result of this is to demonstrate how a cognitive agent is a transducer that maps an input stream of observations to other informational variables specified by an agent's organization and design—those environmental patterns to which it is sensitive. To do this, we frame the notion of a cognitive agent in terms

of computational mechanics—the information theory of structured stochastic processes.

Moreover, there is a practical motivation. Working with processes, as we have above—nominally, infinite sets of infinite sequences and their probabilities—is cumbersome, at best. Typically, we do not care to estimate informations over distributions of infinite pasts and futures. Nor are agents infinitely resourced for doing so.

So, we turn to finitely-specified representations. That is, we turn to models specified by *computational mechanics*—called ϵ-machines and ϵ-transducers—which have especially desirable properties [27, 57] for determining and estimating information processing, both asymptotic properties and, as we show now, moment-by-moment, online statistics. Through their optimal representations of stochastic processes and communication channels, they allow us and cognitive agents, for that matter, to work with finite objects rather than semi-infinite sets and sequences. Helpfully, this often leads to exact, closed-form expressions for many properties of interest [58]. Again, the following assumes that a cognitive agent's internal model, unless otherwise stated, effectively employs its minimal optimal predictor—the observed process' ϵ-machine.

A. Predictive States

We formalize a prediction as a distribution $\Pr(\overrightarrow{X}\,|\,\overleftarrow{x})$ over futures $\{\overrightarrow{X}\}$ with knowledge of a given past $\overleftarrow{x}$. We wish to construct a minimal model that produces optimal predictions for a stochastic process X. Computational mechanics [27] solved this problem in the form of the *ϵ-machine*—a model whose states are the classes defined by an equivalence relation $\overleftarrow{x} \sim \overleftarrow{x}'$ that groups all pasts giving rise to the same prediction $\Pr(\overrightarrow{X}\,|\,\overleftarrow{x})$. These classes are called the *causal states*.

Computational mechanics also provides an analogous solution for optimal predictions of one process, call it X, about another, call it Y [57]. The minimal, optimal model for the conditional input-output process or communication channel—denoted $Y|X$—is called an ϵ-transducer. Since the following makes minimal calls on the latter, here we concentrate on ϵ-machines. Sequels rely more heavily on transducers.

Definition 3. The *causal states of a process* are the members of the range of the function:

$$\begin{aligned}\epsilon\,[\,\overleftarrow{x}\,] = \Big\{ \overleftarrow{x}' \,|\, &\Pr\left(\overrightarrow{X} = \overrightarrow{x}\,|\,\overleftarrow{X} = \overleftarrow{x}\right) \\ &= \Pr\left(\overrightarrow{X} = \overrightarrow{x}\,|\,\overleftarrow{X} = \overleftarrow{x}'\right) \\ &\quad \text{for all } \overleftarrow{x} \in \overleftarrow{X}, \overleftarrow{x}' \in \overleftarrow{X} \Big\}\end{aligned}$$

that maps from pasts to sets of pasts. The set of causal states is denoted $\boldsymbol{\mathcal{S}}$, with corresponding random variable $\mathcal{S}$ and realizations $\sigma \in \boldsymbol{\mathcal{S}}$.

The causal states can be both empirically and measure-theoretically grounded. First, from the (past,future) pairs in a realization $\overleftarrow{x}\,\overrightarrow{x}$ construct an empirical conditional distribution:

$$\widehat{P}_{x_1 \ldots x_n | x_{-k+1} \ldots x_0} = \frac{C_{x_{-k+1} \ldots x_n}}{C_{x_{-k+1} \ldots x_0}} \ , \tag{25}$$

where C_w is the number of times the word w appears in the sequence $x_1 \ldots x_L$. Given sufficient data, it would be desirable to take pasts of arbitrary length and converge towards a prediction conditioned on the *infinite* past:

$$P_{x_1 \ldots x_n | \overleftarrow{x}} = \lim_{k \to \infty} \widehat{P}_{x_1 \ldots x_n | x_{-k} \ldots x_0} \tag{26}$$

with the infinite sequence $\overleftarrow{x} = (\ldots, x_{-1}, x_0)$ of observations stretching into the past.

Second, formally, the conditional predictions $P_{x_1 \ldots x_n | \overleftarrow{x}}$ for all forecast lengths n together describe a *probability measure* over future sequences $\overrightarrow{x} = (x_1, x_2, \ldots)$. In short, a causal state is a measure over future sequences conditioned on past sequences: $\Pr_\mu(\overrightarrow{X}\,|\,\overleftarrow{X} = \ldots x_{-2} x_{-1} x_0)$. We denote it simply $P_{\overleftarrow{x}}$.

Reference [41] established a number of useful convergence and topological properties of ϵ-machine casual states, including:

1. In the language of measures, as an agent collects more observations, $P_{\overleftarrow{x}}$ converges *in distribution*, with respect to the *product topology* of the space of sequences $\mathcal{X}^{\mathbb{N}}$ [41].
2. This also holds for convergence over word distributions $\Pr(x_{1:n})$.
3. A conditional measure is a ratio of likelihoods, as in Eq. (25).
4. For all measures μ on $\mathcal{X}^{\mathbb{Z}}$ and μ-almost every past $\overleftarrow{X} \in \mathcal{X}^{\mathbb{Z}}$, the measures $\eta_\ell[\,\overleftarrow{X}\,]$ defined by:
$$\eta_\ell[\,\overleftarrow{X}\,](U_{0,\omega}) = \Pr_\mu(\omega | x_{-\ell+1} \ldots x_0)$$
converge in distribution to the measure $\epsilon[\,\overleftarrow{X}\,]$: $\eta_\ell[\,\overleftarrow{X}\,] \to \epsilon[\,\overleftarrow{X}\,]$, as $\ell \to \infty$.
5. $\epsilon[\,\overleftarrow{X}\,]$ is the *predictive state* of $\overleftarrow{X}$ and the function that maps to measures over future sequences—$\epsilon[\,\overleftarrow{X}\,] \to \mathbb{M}(\mathcal{X}^{\mathbb{Z}})$—is the *prediction mapping*.

Anticipating future uses, the *causal states of an input-output process*—or *communication channel*—are similarly defined by equivalence classes of channel input sequences and output sequences. The corresponding mapping from an input process to an output process is called the *ϵ-transducer*. The following only considers ϵ-machines for processes, leaving the ϵ-transducer development for channels to a sequel in which they play a pivotal role.

In this way, causal states partition the space of all pasts into sets that are predictively equivalent. The causal state set $\boldsymbol{\mathcal{S}}$ may be finite, fractal, or continuous, depending on X's properties [59].

The following focuses on processes with finite causal-state sets: $|\boldsymbol{\mathcal{S}}| < \infty$. Sequels remove this restriction.

B. Causal-State Processes

Given a stochastic process X, consider a realization:

$$\ldots, x_0, x_1, x_2, \ldots \ .$$

From this, form the associated time series of semi-infinite histories:

$$\ldots, \overleftarrow{x}_0, \overleftarrow{x}_1, \overleftarrow{x}_2, \ldots \ ,$$

where $\overleftarrow{x}_t = \ldots x_{t-2}, x_{t-1}, x_t$.

Now, we use *causal-state filtering* to obtain the causal-state realization—mapping from a time series of observations to a process' causal states:

$$\ldots, \epsilon(\overleftarrow{x}_0), \epsilon(\overleftarrow{x}_1), \epsilon(\overleftarrow{x}_2), \ldots$$
$$\ldots, \sigma_0, \sigma_1, \sigma_2, \ldots \ ;$$

that is, $\sigma_t = \epsilon(\overleftarrow{x}_t)$.

We define two kinds of causal-state behavior of interest:

- The *causal-state process* $\mathcal{S} = \{\mathcal{S}_t : t \in \mathbb{Z}\}$ is the *temporal sequence of causal states* $\mathcal{S}_t = \epsilon[\tau^t \overleftarrow{x}]$; and
- *Recurrent causal-state process* is that series of states after the agent is synchronized to the environment.

We work with the latter, though sequels lay out several of the challenges of working with the transient, nonstationary statistics of the former.

C. ϵ-Machines

The dynamic over the casual states is inherited from the shift operator τ on the process. State-to-state transitions occur when observing a new symbol x, which is appended to the observed history: $\overleftarrow{x} \to \overleftarrow{x}x$. The causal state transition is therefore from $\epsilon[\overleftarrow{x}] = \sigma_i \to \epsilon[\overleftarrow{x}x] = \sigma_j$, and occurs with probability $\Pr(x = x \mid \mathcal{S} = \sigma_i)$.

ϵ-Machines are guaranteed to be optimally predictive because knowledge of what causal state a process is in at any time is equivalent to knowledge of the entire past: $\Pr(\overrightarrow{X} \mid \mathcal{S}) = \Pr(\overrightarrow{X} \mid \overleftarrow{X})$. They are also Markovian in that they render the past and future statistically independent: $\Pr(\overleftarrow{X}, \overrightarrow{X} \mid \mathcal{S}) = \Pr(\overleftarrow{X} \mid \mathcal{S}) \Pr(\overrightarrow{X} \mid \mathcal{S})$. We call these properties together *causal shielding*. ϵ-Machines also have a property called *unifilarity*, which means that knowledge of the current causal state and the next symbol is sufficient to determine the next state. That is to say, $\mathfrak{I}[\mathcal{S}_{t+1} \mid X_t, \mathcal{S}_t] = 0$.

Definition 4. The *ϵ-machine* M_ϵ of a finitary process consists of:

1. Finite *alphabet* $\mathcal{X}$ of k symbols $x \in \mathcal{X}$;
2. *Casual state* set $\boldsymbol{\mathcal{S}} = \{\sigma_1, \sigma_2, \ldots\}$ that consists of transient states whose probabilities vanish and *recurrent* states $\tilde{\sigma}$ whose probability converges to a constant $\Pr(\tilde{\sigma}) > 0$. And;
3. *Causal dynamic*—set of k (possibly infinite dimension) symbol-labeled transition matrices $T^{(x)}$, $x \in \mathcal{X}$: $T^{(x)}_{ij} = \Pr(\sigma_j, x \mid \sigma_i)$.

This defines its own class of optimal representations for a wide range of stochastic processes. That said, we can draw several parallels with existing classes of process representations. The definition here identifies an ϵ-machine as a *hidden Markov model* (HMM). Not all HMMs are ϵ-machines. However, the ϵ-machines here are. An ϵ-machine may be graphically shown as an HMM with a directed graph where the causal states are depicted by vertices and transitions between them by directed edges labeled with the symbol emitted on transition followed by the probability of transition; e.g., $x : \Pr(x)$. The time indexing is as follows: if at time t, an ϵ-machine is in state $\mathcal{S}_t$, it emits symbol x_t and transitions forward to the next state $\mathcal{S}_{t+1}$. Notice that due to unifilarity, there is at most one transition from each causal state per symbol.

D. Agent-Environment Synchronization

Proposition 4. *Given the causal state at time $t-1$, the causal state at time t is independent of the causal states at earlier times $t < t-1$.*

Proof. See Lemma 6 (ϵ-Machines are Markovian) in Ref. [27]. □

That is, the causal-state process is order-1 Markov:

$$\Pr(\sigma_t | \ldots \sigma_{t-2}\sigma_{t-1}) = \Pr(\sigma_t | \sigma_{t-1}) \ .$$

Markovity induced by the causal states makes it particularly straightforward to give the recurrent causal-state process directly in terms of its word distributions:

$$\Pr(\widetilde{\sigma}_{0:n}) = (\pi_0)_{\widetilde{\sigma}_0} \prod_{t=0}^{n} T_{\widetilde{\sigma}_t, \widetilde{\sigma}_{t+1}} \ ,$$

where $\widetilde{\sigma}_t$ is a recurrent causal state, T is the state-transition operator, and $(\pi_0)_{\widetilde{\sigma}_0}$ is the initial causal state distribution. (The tildes identify recurrent states.)

The simplicity of the casual-state process demonstrates concretely why the causal states are so useful. For one, the summary they give of the past $\overleftarrow{X}$ is sorely needed to make various potentially-infinite calculations tractable.

Leveraging this, we define *agent-environment synchronization* when the agent is fully tracking the environment. First, define a synchronizing word $\widehat{w}$: $\Pr(\widetilde{\sigma}|\widehat{w}) = 1$, for one $\widetilde{\sigma} \in \boldsymbol{\mathcal{S}}$, and where $\widehat{w}$ is one of the words $\mathcal{L}(X)$ generated by the process: $\widehat{w} \in \mathcal{L}(X)$. The informational definition of synchronization is that the agent exactly knows the current state: $\mathfrak{I}[\sigma_t \mid \widehat{w}] = 0$.

E. Agent Operation

There are two principle modes of operation for ϵ-machines: Recognizing sequences of environment behaviors versus generating environment control signals:

- *Recognition mode*: This addresses the question, is a given word $w \in \mathcal{X}^*$ in the set (or language) recognized by the ϵ-machine? Is $w \in \mathcal{L}(M)$?
- *Generation mode*: The mode concerns what words $w \in \mathcal{X}^*$ are emitted by an ϵ-machine.

As noted below, these modes of operation come into play when an agent encounters stimuli disallowed by its internal model. That is, when it resets its state in response.

Before use, an ϵ-machine must be properly initialized. In either recognition or generation mode, the procedure to initialize an ϵ-machine sets the initial causal-state distribution: Set the current state to ϵ-machine's unique start state $\mathcal{S} = \sigma_0$ and set the current state distribution π_t to have unity probability on σ_0: $\pi_0 = (1, 0, 0, \ldots)$.

Given an ϵ-machine, its asymptotic causal-state distribution $\widehat{\pi}$ is the left-eigenvector of the causal-state transition operator $T = \sum_{x \in \mathcal{X}} T^{(x)}$:

$$\widehat{\pi} = \widehat{\pi} T \ ,$$

normalized in probability: $\sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) = 1$.

In each mode, the ϵ-machine operates step-by-step, sequentially reading the symbols in a word, to accept or reject the word as follows.

- *Acceptance*: Reach the input word's last symbol while in recurrent causal state $\sigma_t \in \boldsymbol{\mathcal{S}}$.

  Sequentially reading symbols from the input, update the current state to that reached by transition labeled by x_t and update current state distribution:

  $$\pi_{t+1}(x) = \frac{\pi_t T^{(x)}}{\sum_x \pi_t T^{(x)}} \ .$$

- *Rejection*: While reading symbols from the input the agent encounters a disallowed symbol; that is, from the current state there is no transition labeled with that symbol.

  Then, reset the current state to the agent's unique start state σ_0 and reset the current-state distribution π_t to unit probability on the start state σ_0:

  $$\pi_{t+1} = (1, 0, 0, 0, \ldots) \ .$$

  For generation mode the ϵ-machine operates according to a companion *emission* procedure that simply follows the allowed transitions emitting the symbols labeling each.

The probability of generating word $w = x_0 x_1 \ldots x_{\ell-1}$ given start distribution π_0 is:

$$\begin{aligned} \Pr(w) &= \pi_0 \prod_{i=0}^{\ell-1} T^{(x_i)} \\ &= \pi_0 T^{(\omega)} \ . \end{aligned} \tag{27}$$

If an ϵ-machine starts with $\widehat{\pi}$, then $\pi_0 = (1, 0, ...)$. Recall that the start state corresponds to the asymptotic state distribution; a condition representing unbiased information about the environment's state.

Proposition 5. *If an ϵ-machine used $\pi_0 = \widehat{\pi}$ to generate word w, then $\Pr(w)$ is w's asymptotic stationary probability.*

Proof. By elementary Markov chain theory [44]. □

Definition 5. If M is an ϵ-machine, its *output process* $X(M)$ consists of the word distributions given by Eq. (27) above.

Proposition 6. *If starting with $\widehat{\pi}$, an ϵ-machine's output process is stationary.*

Proof. Elementary Markov chain theory [44] shows that word distributions are time-shift invariant. □

F. Intrinsic Information Processing

There are infinitely many possible optimal predictive models—if thinking of state-based models, imagine adding redundant or even useless states. Helpfully, computational mechanics established that the ϵ-machine is a process' minimal and unique predictive model in the sense that the amount of information stored by the causal states is smaller than (or equal) to any other possible *prescient* (equally predictive) rival. We quantify this via the Shannon entropy of the causal-state distribution—this is the *statistical complexity*: $C_\mu = \mathfrak{I}[\mathcal{S}] = \mathfrak{I}[\widehat{\pi}]$.

With this, there are three basic informational invariants that describe a stationary process: h_μ is a process' rate of information creation, C_μ is the amount of history a process remembers from its past $\overleftarrow{X}$—the historical information stored in the causal states $\boldsymbol{\mathcal{S}}$; and $\boldsymbol{E}$ is the amount of that stored information communicated to the future $\overrightarrow{X}$.

Generally, the following refers to an environment's *inherent information* if an optimally-predicting agent uses the environmental process' ϵ-machine and is synchronized.

If the agent uses a model other than the ϵ-machine, then the information available is relative to that model and to being synchronized or not. Hence, this highlights the distinction between "subjective" self-information—using any model—and "inherent" self-information—using the ϵ-machine. Section VI D below illustrates the consequences of an agent using an incorrect environment model.

Finally, we generically refer to the time series of temporal information measures and the causal-state and prediction processes collectively as *information processes*. These are the time series generated during an agent's online operation that allow us to monitor and diagnose the informational processing that it performs while interacting with an environment.

VI. INTRINSIC SEMANTICS OF INFORMATION PROCESSES

Generally, one concludes that an ϵ-machine captures the patterns in a stochastic process [27]. However, focusing on the global, time-asymptotic view side-steps a key question that Fig. 3 highlights:

What does a particular measurement mean?

To address this, the following reviews and then applies the notion introduced some time ago in Ref. [59] of *measurement semantics*. This refers to the (informational) meaning revealed to an agent via measurements of its environment with respect to its internal model of that environment.

Said more plainly, we assume an agent has a model in its "head" that it uses to interpret the incoming sequence of measurements. The following explores the relationship between incoming "raw" observations and how an agent deploys a (good or bad) model of the observed process' generator to interpret them.

It might use this model to "understand" patterns in the process or simply may use it to predict future observations. We start with the latter first.

Two points to clarify first. The following assumes an agent has a model of its environment. As such, the following does not discuss how an agent obtains its model. This certainly brings up important, but for now peripheral, issues; for example, the issues of statistical estimation, inference, and overfitting. See, for example, Ref. [60] and references therein. Beyond inference, though, there are many competing ways an agent can have a model of its environment. It can, for example, simply sample a given model distribution—a topic for a separate development.

In addition, the following assumes the agent uses the incoming observations and its model to synchronize to the environment's state. The way in which it does this can be quite challenging. It is its own topic; see Ref. [61].

A. Prediction Semantics

One interpretational setting is to use past observations $\overleftarrow{x}_t$ up to time t to predict the future, say, x_{t+1} or even the whole future $\overrightarrow{x}_t = x_{t+1}, x_{t+2}, \ldots$. But to what end? Specifically, even mere prediction begs asking, What is the meaning of the particular measurement $x_{t+1} \in \mathcal{X}$?

Harking back to Sec. IV, we have the following.

Definition 6. Given a particular past $\overleftarrow{x}_t \in \overleftarrow{X}$, recall that Shannon defined the amount of *self-information* an agent gains in observing $x \in \mathcal{X}$ to be [45]:

$$-\log_2 \Pr(X_t = x) \ . \tag{28}$$

This is, in short, an agent's degree of surprise on observing x_{t+1}. As Shannon formulated it and as noted above, the degree of surprise is the most basic concept in information theory. Beyond the intuition motivating self-information, Ref. [45] did not specify exactly how one obtains the probability $\Pr(X_t = x)$. It is assumed to be available to an agent ... (or, in Shannon's setting, available to the communications channel analyst.)

And so, we must refer to this as the *subjective self-information* since, nothing else said, it does not specify how the probability is determined. (This is in stark contrast to the comments at the beginning, which assumed probabilities were given.) Specifically, any interpretive model could be used to develop $\Pr(X_t = x)$ and the self-information would change accordingly. The need to address this issue is now clear. The following assembles the required components.

We can specialize the above definition of *subjective* self-information by constraining the model used, rather than it being arbitrary. (Again, that arbitrariness led us to the label "subjective".) Having identified this relativity, when using information theory one is chastened to specify at the outset the model class of process generators with which one works.

Heeding this, one can then specify a model class, such as Markov chains or hidden semi-Markov processes, as appropriate. However, there is still model choice within a given class and so follow-on informational interpretations are subjective. Computational mechanics' introduction of a process' ϵ-machine solves this problem, grounding interpretation in a process' minimal and unique representation that comes from the process itself. Seemingly simple, this is one of computational mechanics' broader contributions as it removes unnecessary subjectivity in applying information theory to a given stochastic process..

Computational mechanics adopts a companion view of a communication channel that refines Shannon's notion of the channel *receiver* by making explicit the observing entity—the agent; in fact, one that is optimally predicting. In short, for objective or intrinsic information measures, an agent employs a process' ϵ-machine as its internal model.

A key consequence of its optimality is that using the ϵ-machine as the agent's internal model grounds much of information theory that is unspecified or, as noted, subjective.

Definition 7. Given a stationary process X and an agent synchronized to it that uses X's ϵ-machine $M = \{\mathcal{X}, \boldsymbol{\mathcal{S}}, \{T^{(x)}_{\mathcal{S}\to\mathcal{S}'} : x \in \mathcal{X}\}, \mu_0\}$, the *intrinsic self-*

information in observing x is:

$$\begin{aligned} &-\log_2 \Pr(X_t = x) \\ &\quad = -\log_2 \Pr(\overleftarrow{x}_t) \Pr\left(\sigma_t \rightarrow_x \sigma_{t+1} | \overleftarrow{x}_t\right) \ . \end{aligned} \tag{29}$$

Here, the self-probability $\Pr\left(\sigma_t \rightarrow_x \sigma_{t+1}\right)$ is calculated from the ϵ-machine, as follows: (i) Note the probability $\Pr(\sigma_t | \overleftarrow{x}_t)$ of being in state σ having observed $\overleftarrow{x}_t$. Applying the $\epsilon(\cdot)$, this is unity. And, (ii) multiply that by the (transition) probability $T^{(x)}_{\mathcal{S} \rightarrow \mathcal{S}'}$ of taking the x-labeled transition. That is:

$$\Pr(X_t = x) = \Pr(\overleftarrow{x}_t) T^{(x_t)}_{\epsilon(\overleftarrow{x}_t) \rightarrow_x \mathcal{S}'} \tag{30}$$

$$= \Pr(\overleftarrow{x}_t) T^{(x)}_{\sigma_t \rightarrow_x \mathcal{S}'} \ . \tag{31}$$

Note the dependence on the process' asymptotic invariant measure μ, implicitly in using $\Pr(\overleftarrow{x}_t)$.

Again, using the ϵ-machine is a natural choice given that it is the process' minimal optimal predictor. And so, follow-on results inherit a number of desirable properties—such as, actually describing the given process and facilitating efficient estimation of informational properties. In addition, Ref. [27] established that the causal-state process—the temporal sequence of causal states the agent visits—is an order-1 Markov process. This greatly simplifies much technical development since each causal state summarizes in a single RV $\sigma_t = \epsilon(\overleftarrow{x}_t)$ all of the prediction-relevant information from the past.

To emphasize, the present setting and the following assume the agent is *synchronized* to the incoming process—it knows which causal state (σ above) the process generator is in. For stochastic processes generated using finite-memory there are two kinds [61]: (i) exact synchronization in which the agent knows the state after a finite series of observations and (ii) asymptotic synchronization where an infinite series of observations is required and the agent converges (exponentially fast) to synchronization.

Agent-environment synchronization is an important property for the results here. A number of consequences arise when removing the synchronization requirement, especially when addressing the potentially transient, nonstationary epoch prior to synchronization. Fortunately, the conditions of exact and asymptotic synchronization can be substantially broadened to include processes generated by *path-mergeable* HMMs [61].

B. Online Prediction: An Example

The following assumes the agent uses the process' minimal optimally-predictive model—the ϵ-machine.

To illustrate the operation of prediction, let's consider how an agent sequentially monitors (optimally) observations in a realization of the Even Process:

t	$=0$	1	2	3	4	5	6	7	8	9	10	11
s_t	$=0$	1	1	1	1	0	1	1	0	1	1	1

Now, at time $t = 11$, the agent measures $s_{11} = 1$. (For reference, see the Even Process' ϵ-machine in Fig. 8(d).)

How much information does $X_{11} = 1$ convey to the agent? We apply the self-information conditioned on the preceding eleven observations:

$$\mathfrak{I}\left[s_{11} \mid s_{10} = 1, s_9 = 1, \ldots\right] \approx h_\mu \tag{32}$$

$$\approx 0.585 \text{ bits} \ . \tag{33}$$

This is the degree of the observer's surprise (unpredictability).

C. Measurement Semantics

Crucially, this indicates uncertainty but does not determine what the event $s_{11} = 1$ means to the observer!

We have one event—$s_{11} = 1$, but there are two contexts or interpretive levels in which to interpret it. The first is how much it is unanticipated—the semantics of prediction. The second, concurrent setting is how it updates the agent's actual anticipation. That is, to what model state σ_t does the observation take the observer? The meaning to the agent comes from a tension between representation of the same event $s_{11} = 1$, but at different levels or contexts. Here, we have:

1. Level 1 is the data stream and the event is a measurement;
2. Level 2 is the agent and the event updates its model.

With this in mind, we have the following definition of an agent's observational semantics.

Definition 8. The *degree of meaning* of observing $x \in \mathcal{X}$:

$$\Theta(x) = -\log_2 \Pr(\rightarrow_x \sigma) \ , \tag{34}$$

where the arrow notation signifies $\sigma \in \boldsymbol{\mathcal{S}}$ is the causal state *to which* x brings the agent.

Said differently, this broadens our understanding of causal states—they are *contexts of interpretation* within which observed values appear. The set $\boldsymbol{\mathcal{S}}$, thus, is agent's palette of semantic interpretations of environment behaviors.

In the preceding example of the Even Process, the state selected is D which has asymptotic probability $\Pr(\mathcal{S} = D) = 2/3$) and so:

$$\begin{aligned} \Theta(x_{11} = 1) &= -\log_2 \Pr(\sigma_{11} = D) \\ &= -\log_2(2/3) \\ &\approx 0.58496 \text{ bits} \ , \end{aligned}$$

of semantic information.

Moreover:

Definition 9. The *meaning content* is the state σ selected from the palette of anticipations—the model's state set $\boldsymbol{\mathcal{S}}$.

Again, in the preceding example of the Even Process, $x_{11} = 1$ selects state D. And this means that the 1—compared to other 1s that may be emitted—is the second 1 in the symbol-pairing—$0(11)^n0$—that the Even Process emits.

And, this highlights the origin of the measurement's meaning *within* the set of the agent's causal states. That is, the meaning depends on the entire set of interpretive contexts—within the causal state set $\boldsymbol{\mathcal{S}}$. That is to say that the interpretive context of an incoming observation is the causal state the agent is currently in. For one, the causal state has an associate prediction—its future morph. For another, the current causal state determines which next state the agent will move to. Both predicting the next observation and the next state are interpretations of what the current observation means to the agent.

D. Meaninglessness

It is important to highlight that this measurement semantics allows for measurements to be meaningless. Here, are several cases.

1. Given how the ϵ-machine operates—recall the acceptance and rejection modes described above—one specifies an initial state distribution. And, under one operation mode, the ϵ-machine puts all of the probability on the start state. That is, π_t is set to $\pi_0 = (1, 0, 0, 0, \ldots)$. Or, in other words, $\Pr(\sigma_0) = 1$. And, the symbol observed is $x_0 = \lambda$. The degree of meaning, then, is:

$$\begin{aligned} \Theta(x_0 = \lambda) &= -\log_2 \Pr(\sigma_0) \\ &= -\log_2 1 \\ &= 0 \ . \end{aligned}$$

2. Disallowed transition: Upon seeing a disallowed symbol, the agent resets its ϵ-machine to the start state—the state of total ignorance. As just noted, this means the state, being the start state, has full probability and so is meaningless. This is intuitive: The disallowed transition or forbidden symbol observed is meaningless. The model has no interpretation to give—nothing to say: $\Theta = 0$.

Thus, by the above, the start state is meaningless. Indeed, having seen nothing, all futures are possible.

Somewhat glibly, meaningless measurements are (very!) informative:

$$\begin{aligned} -\log_2 \Pr(\sigma_t \rightarrow_{x_t} \mathcal{S}_0) &= -\log_2 0 \\ &= \infty \ . \end{aligned}$$

A meaningless measurement is informative in the sense that a zero probability observation is infinitely surprising. And, moreover, a meaningless measurement implies the existence of (at least one) zero-probability measurement.

Thus, even if the agent has an incorrect model of a process there is a semantics of its (likely misleading) interpretations of the environment behaviors. In such circumstances, there can be many symbols and transitions that the agent interprets as disallowed. No matter, the above theory properly describes the semantics according to the agent. Section VII E illustrates the circumstance when an agent uses as its internal model the Even Process ϵ-machine to (incorrectly) interpret an environment obeying the Golden Mean Process.

Finally, how this semantic theory applies to hierarchically-structured processes—such as an environment at the onset of chaos or the particle interactions in cellular automata distributed computation—are the subjects of a sequel.

E. Total Semantic Information

It turns that the total amount of semantic information in a process is related to its information storage.

Theorem 1. *A process' total average semantic information is its statistical complexity:*

$$\langle \Theta(x) \rangle = C_\mu \ .$$

Proof:

$$\begin{aligned} \langle \Theta(x) \rangle &= \sum_{\sigma \in \boldsymbol{\mathcal{S}}} \Pr(\sigma) \Theta(x) \\ &= -\sum \Pr(\sigma) \log_2 \Pr(\sigma) \\ &= \mathfrak{I}\left[\mathcal{S}\right] \\ &= C_\mu \ . \end{aligned}$$

In other words, the average amount of meaning is the statistical complexity C_μ. This gives an insightful connection between a process' internal structure and the semantics an observer attributes to observations.

Tables IV, V, VI, and VII present an agent's informational and semantic analysis of Sec. VII's example processes, respectively.

F. Ergodic Theory Redux

The following establishes that the information processes generated by a (finitary) agent are well-behaved in the sense that the information processes resulting from observing a stationary and ergodic environment are well-behaved. Note that such properties are an aid to further, downstream processing that an agent is tasked to perform.

This holds for any class of generated information process—whether entropy rate or excess entropy or others based on the various self-informations or densities

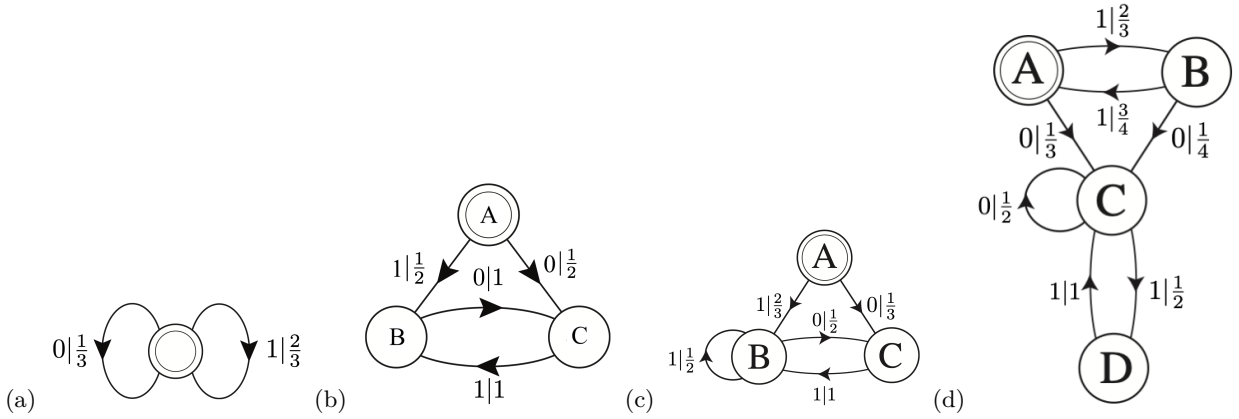

FIG. 8. Example ϵ-machines for: (a) Biased Coin; (b) Period-2; (c) Golden Mean; and (d) Even Processes. Causal states are circles, with the start state having an inscribed circle. State-to-state transitions are labeled $x|p$, where $x \in \mathcal{X}$ is the emitted symbol and $p \in [0, 1]$ is the transition probability.

(self-informations) developed above. Given this, we then develop analogous results for when an agent observes an ergodic process in its environment.

Proposition 7. *If X is stationary, then the prediction mapping $\epsilon_t = \epsilon(\overleftarrow{X}_t)$ is a random variable that is stationary.*

Proof. ϵ_t is a first-order Markov process [27]. □

The *prediction process* $X_\epsilon = \{X_t|\epsilon_t : t \in \mathbb{Z}\}$ is the stochastic process of predictions $\Pr(\overrightarrow{X}|\cdot)$.

Proposition 8. *The prediction process is stationary and ergodic.*

Proof. ϵ_t is a first-order Markov process [27]. □

Similarly, the causal-state process is statistically well-behaved.

Proposition 9. *If X is stationary and ergodic, then the causal state process $\{\mathcal{S}_t : t \in \mathbb{Z}\}$ is stationary and ergodic.*

Proof. The causal-state process is a finite-range function of the process. And so, by previous propositions, since the process is stationary and ergodic, the causal-state process is stationary and ergodic. □

Define the *prediction uncertainty process*: $h_\mu(t) = \mathfrak{i}[x_t \mid \overleftarrow{x}_t]$. It is also given by: $h_\mu(t) = \mathfrak{i}[x_t \mid \mathcal{S}_t]$. The latter expression is often a direct and efficient way to generate the prediction process, if an ϵ-machine is in hand.

Proposition 10. *The prediction uncertainty process is stationary and ergodic.*

Proof. The prediction uncertainty process is a finite-range function of the process. And so, by previous propositions, since it is stationary and ergodic. □

Define the *causal-state uncertainty process*: $C_\mu(t) = \mathfrak{i}[\sigma_t = \epsilon(\overleftarrow{x}_t)]$. This too is stationary and ergodic if X is:

Proposition 11. *The causal-state uncertainty process is stationary and ergodic.*

Proof. The causal-state uncertainty process is a finite-range function of the process. And so, by previous propositions, since the latter is stationary and ergodic, it is too. □

Having seen the pattern in the preceding results, we generalize. Let $\mathfrak{I}(\cdot)$ be any function defining a temporal information atom—as those laid out above in Table III. The preceding observations generalize to establish that functions—in particular, the self-informations $\mathfrak{I}(X)$—of stationary, ergodic processes are stationary and ergodic.

Proposition 12. *Self-information processes $\mathfrak{I}\left[A \mid \overline{A}\right](t)$ are stationary and ergodic, if the process is.*

Proof. The self-information processes are finite-range functions of the process. And so, by previous propositions, they stationary and ergodic. □

This is all to say that an agent has well-defined and statistically well-behaved informational quantities to work with in its more sophisticated downstream cognitive processing, such as further-derived information-theoretic quantities, in making inferences about them, in quantitatively monitoring its inferences, in identifying fluctuations, and in decision-making and taking actions on the environment.

VII. EXAMPLE INFORMATION PROCESSES

The following analyzes concrete examples of information processes produced an agent observing an environment governed by finite-memory generators: (i) wholly unpredictable, (ii) predictable, (ii) structured, but finite Markov

Observer's Analysis of Biased Coin Process				
State	Measurement x	Surprise $-\log_2 \Pr(x)$ [bits]	Semantic State σ: Meaning	Degree of Meaning $\Theta(x)$ [bits]
A	λ	Not Defined	No Measurement	0.0
A	1	0.585	A: Sync	0.0
A	0	1.585	A: Sync	0.0

TABLE IV. Information and Semantic Analysis of the Biased Coin Process, with 1s-bias of $\Pr(x = 1) = 2/3$. Recall Fig. 8(a).

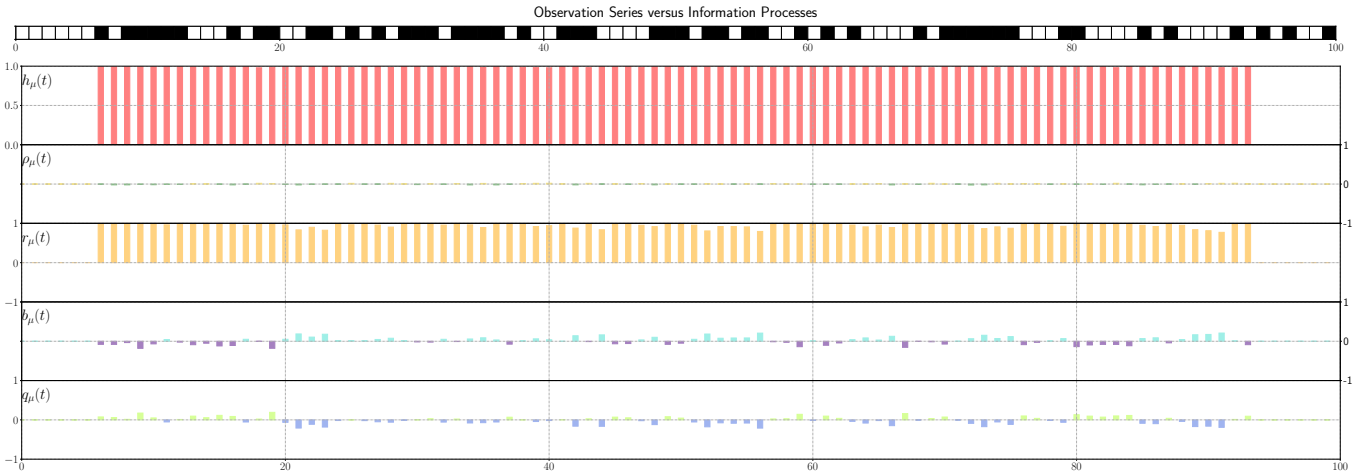

FIG. 9. Biased Coin Process observed time series realization (top) versus several of its information processes (below): entropy rate $h_\mu(t)$, anticipated information rate $\rho_\mu(t)$, ephemeral rate $r_\mu(t)$, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$. The information process values for the first and last 6 steps are not shown, due to their being estimated using a window length of 13: 6 steps for the history and for the future words, plus the present (1 step). This holds for all of the following information process time series.

order, and (iv) structured and infinite Markov order. In other words, several relatively low-complexity (easy to explain) environments that range from predictable memoryless to unpredictable infinite-range dependencies—a suite that illustrates the breadth of basic results of interest. See Fig. 8 for their state transition diagrams.

Building on these and the general development of information processes, the following section establishes that, whatever its role, an agent has access to and generates well-behaved information processes—first, such derived processes are stationary and second they are ergodic. Tables IV, V, VI, and VII present an agent's informational and semantic analysis of the example processes, respectively. And, Figs. 9, 10, 11, and 12 plot typical realizations of the examples' information processes.

References [24, 40] provide fuller informational analyses for these and other example processes. Here, we simply focus on aspects related to a cognitive agent's associated information processes.

A. Unpredictable: Independent Identically-Distributed

The paradigm of a random process is the fair (or biased) coin—one in the family of *independent identically distributed* (IID) processes—coins, dies, and the like. Consider the particular case of a process of identical independently distributed variables; i.e., a coin flip with probability $\{\Pr(H) = p, \Pr(T) = 1 - p\}$ repeated infinitely many times.

Figure 8(a) gives the ϵ-machine state-transition diagram. For all processes in the IID family, there is only a single causal state $\boldsymbol{\mathcal{S}} = \{A\}$. For the binary alphabet process generated the state-to-state transitions are given by two 1×1 symbol-labeled transition matrices: $\{T^{(0)} = (p), T^{(1)} = (1-p)\}, p \in [0, 1]$. The initial measure over states is also trivial: $\pi_0 = \{1\}$. π_t is invariant over time.

Its basic informational quantities are: (i) entropy rate $h_\mu = 1$ bit per time step, (ii) statistical complexity $C_\mu = 0$ bits, and (iii) excess entropy $\boldsymbol{E} = 0$ bits. All as one would expect for a fully random, memoryless process.

Each measurement X_t being independent, the mutual information $\mathfrak{I}[X_t : X_{t'}] = 0$ for all $t \neq t'$. Therefore, all information atoms in Fig. 7(left) vanish except for $r_\mu = h_\mu = \mathfrak{I}[X_t]$ and the infinite amount of information in the past and future.

Table IV presents an agent's informational and semantic analysis of an IID process in terms of the self-information (transition) surprise $-\log_2(x)$ and degree of meaning $\Theta(x)$: IID processes are always synchronized and the

Observer's Analysis of Period-2 Process				
State	Measurement x	Surprise $-\log_2 \Pr(x)$ [bits]	Semantic State σ: Meaning	Degree of Meaning $\Theta(x)$ [bits]
A	λ	Not Defined	No Measurement	0.0
A	1	1.0	B: Sync	1.0
A	0	1.0	C: Sync	1.0
B	1	∞	A: Loose sync; reset	0.0
B	0	0.0	C: Deterministic 1	1.0
C	1	0.0	B: Deterministic 0	1.0
C	0	∞	A: Loose sync; reset	0.0

TABLE V. Information and Semantic Analysis of the Period-2 Process. Recall Fig. 8(b).

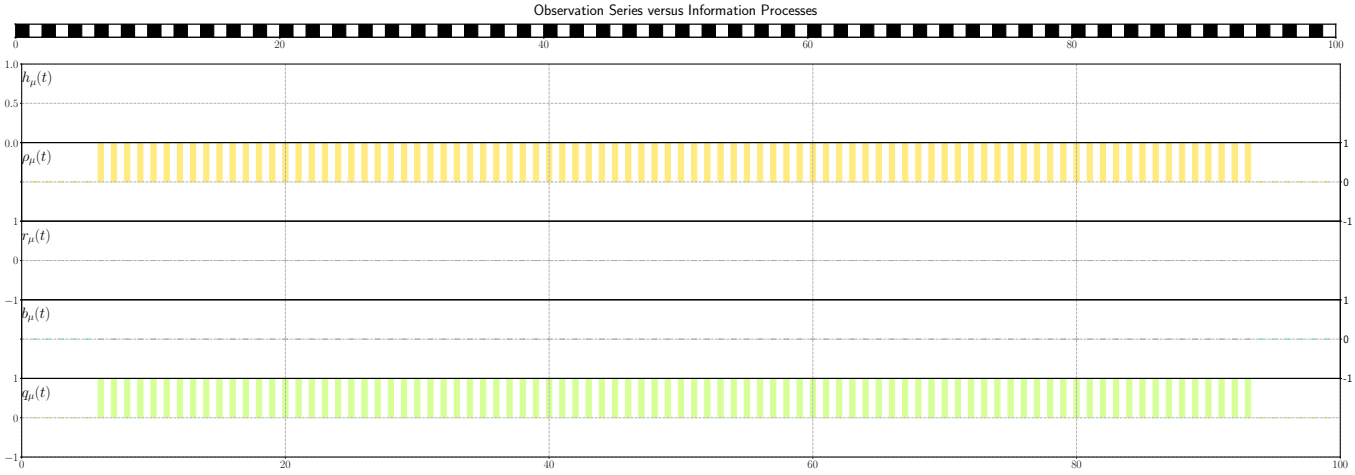

FIG. 10. Period-2 Process observed time series sample versus several of its information processes: entropy rate $h_\mu(t)$, anticipated information rate $\rho_\mu(t)$, ephemeral rate $r_\mu(t)$, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

semantics is meaningless, as one would expect for a structureless process.

Finally, there are the Biased Coin's information processes. Figure 9 presents a realization and its information processes: entropy rate $h_\mu(t)$, anticipated information rate ρ_μ, ephemeral rate r_μ, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$. As expected, every succeeding observation is highly surprising ($h_\mu = 1$ bit) and no information in the future is anticipated ($\rho_\mu = 0$ bits). All of the information h_μ produced at each time step is forgotten ($r_\mu \approx 1$) and none is stored ($b_\mu \approx 0$). The fluctuations seen in these latter information processes reflect their being empirically estimated from a finite time series of observations (length $\ell = 10^6$ symbols.)

B. Predictable: Period-n

Now consider the other extreme of predictability—a completely determined period-2 process $\ldots 010101 \ldots$. Figure 8(b) presents the ϵ-machine that generates a period-2 binary process. There is a single transient state A and the initial distribution is $\pi = \{1, 0, 0\}$. The recurrent states B and C are equally likely—$\pi = \{0, 1/2, 1/2\}$—which state distribution is time invariant.

The state transition matrices for the binary generator are:

$$T^{(0)} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } T^{(1)} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

All conditional entropies vanish, including the information in the past and future, as any single measurement determines the value of X_t for all time indices. The only remaining quantity is q_μ, which contains all the process information. In the case of a period-n process, the value of this atom in bits is exactly $\log_2 n$. Table V gives an agent's informational and semantic analysis of a period-2 process. When disallowed symbols violate the periodicity, the ϵ-machine resets to the (meaningless) start state, having lost the oscillation's phase, while the agent is infinitely surprised that this has occurred.

Finally, there are Period-2's information processes. Figure 10 presents a realization and its information processes: entropy rate $h_\mu(t)$, anticipated information rate ρ_μ, ephemeral rate r_μ, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$. As expected, every succeeding observation is determined ($h_\mu = 0$ bit) and all of this future information is anticipated ($\rho_\mu = 1$ bits). Since no information h_μ produced at each time step there is none to forget ($r_\mu \approx 0$) and none to store stored ($b_\mu \approx 0$).

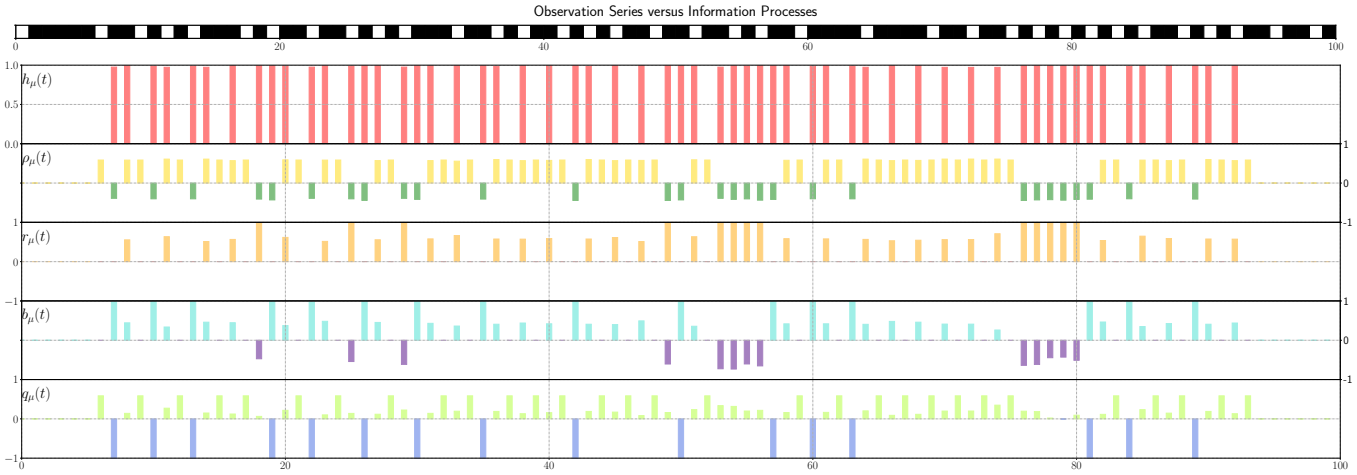

FIG. 11. Golden Mean Process observed time series sample versus several of its information processes: entropy rate $h_\mu(t)$, anticipated information rate $\rho_\mu(t)$, ephemeral rate $r_\mu(t)$, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

Observer's Analysis of Golden Mean Process				
State	Measurement x	Surprise $-\log_2 \Pr(x)$ [bits]	Semantic State σ: Meaning	Degree of Meaning $\Theta(x)$ [bits]
A	λ	Not Defined	No Measurement	0.0
A	1	0.585	B: Sync	0.585
A	0	1.585	C: Sync	1.585
B	1	1.0	B: Random 1	0.585
B	0	1.0	C: Isolated 0	1.585
C	1	0	B: Deterministic 1	0.585
C	0	∞	A: Loose sync; Reset	0.0

TABLE VI. Information and Semantic Analysis of the Golden Mean Process. Recall Fig. 8(c).

There is a single bit ($q_\mu = 1$ bit) of phase information of the period-2 cycle.

The extreme examples of utterly random and perfectly predictable are simple enough to work through, but typically calculating the information dynamics requires taking entropies over infinite sequences, which is not practical. Therefore, we turn to the use of optimal finite models and the theory of computational mechanics.

C. Finite Markov Order: Golden Mean

As one sees from Ref. [26]'s survey (Fig. 13 there), the vast majority of structured stochastic processes lie between the two preceding extremes of predictability. As a typical example of these intermediate-complexity process generators consider the Golden Mean Process. See Fig. 8(c) for its ϵ-machine state transition diagram, which generates all binary sequences except for those with consecutive 0s. That is, only the word $w = 00$ is forbidden; otherwise the generated realizations are random. It is easy to see where in the state-transition diagram this restriction arises: When the ϵ-machine is in state C it must transition to state B and emit a 1 with probability 1. A simple way to summarize this and so characterize the Golden Mean Process is to give its list of *irreducible forbidden words* $\mathcal{F} = \{00\}$.

As in the period-2 process, there is a single transient state A and the initial distribution is $\pi_0 = \Pr(A, B, C) = \{1, 0, 0\}$. The state transition matrices for the Golden Mean Process generator are:

$$T^{(0)} = \begin{pmatrix} 0 & 0 & 1/3 \\ 0 & 0 & 1/2 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } T^{(1)} = \begin{pmatrix} 0 & 2/3 & 0 \\ 0 & 1/2 & 0 \\ 0 & 1 & 0 \end{pmatrix} .$$

The recurrent states B and C are visited with probabilities $\widehat{\pi} = \{0, 2/3, 1/3\}$—which state distribution is time invariant after the first time step.

The informational analysis says the entropy rate is $h_\mu = 2/3\ \mathfrak{I}(1/2) = 2/3$ bits per emission. The statistical complexity is $C_\mu = \mathfrak{I}(2/3)$ bits, as is the excess entropy. The generated Golden Mean process is Markov order 1. For these see Ref. [50].

Table VI gives an agent's informational and semantic analysis of the Golden Mean Process.

And, Fig. 11 presents its information processes: entropy rate $h_\mu(t)$, anticipated information rate ρ_μ, ephemeral rate r_μ, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

As concrete examples of the Golden Mean Process' information processes we examine the prediction and statistical complexity processes—$h_\mu(t)$ and $C_\mu(t)$. First, we need the state distribution $p(t)$ as a function of time: this is simply $p(0) = (1, 0, 0)$ and $p(t) = (0, 2/3, 1/3), t = 1, 2, 3, \ldots)$.

From this, we then calculate the time-local quantities from looking at machine states—out of which set of values this or that information process will consist of a time series of samples. That is, for the prediction process we have uncertainties in state A for first time step: for observing a 1 $h_\mu(t = 0, x_0 = 1) = \mathfrak{I}(2/3)$, a 0 $h_\mu(t = 0, x_0 = 0) = \mathfrak{I}(1/3)$. For second step; $h_\mu(t = 1, x_{0:1} = 11) = 1$ bit, $h_\mu(t = 1, x_{0:1} = 10) = 0$ bit. $h_\mu(t = 1, x_{0:1} = 01) = 0$ bit. At this point, having these initial observations, the agent is synchronized to the environment behavior. And so, values in the prediction process are samples of these values, sampled according to the history seen. Then we have the statistical complexity process: $C_\mu(t = 0) = 0$. $C_\mu(t) = \mathfrak{I}(2/3)$, $t = 1, 2, 3, \ldots$. This is quite simple.

Finally, there are Golden Mean's information processes. Figure 11 presents a realization and its information processes: entropy rate $h_\mu(t)$, anticipated information rate ρ_μ, ephemeral rate r_μ, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

Given that there are no consecutive 0s, if a 0 is observed there is no uncertainty in the next observation (it must be a 1). At those times (denote them t), $h_\mu(t) = 0$. Otherwise, (denote those times t'), a fair coin flip is anticipated: $h_\mu(t') = 1$ bit. At times t' the bit of created information is either completely lost ($r_\mu(t') = 1$ bit) or some of it is stored ($b_\mu(t') > 0$) such that $r_\mu + b_\mu = h_\mu$.

We notice there are several negative informations—$b_\mu(t)$, ρ_μ, and q_μ. As explained at the end of Sec. IV D, negative informations are to be expected in the setting of temporal information measures.

Overall, we see how the various information processes make clear that the Golden Mean Process is a complicated combination of the two information processing modes of the Biased Coin and Periodic Processes. There is more to say, of course. What is missing is a more mechanistic explanation of how the information quantities are being stored and processed. This is the burden of a sequel. Here, our goal was to motivate and then explore information processes. The next example illustrates the need for this amply.

D. Infinite Markov Order: Even

The fourth example to consider is more complex still. This is the Even Process, whose ϵ-machine is shown in Fig. 8(d). Whereas the Golden Mean Process was determined by finite-length word restrictions, the Even Process exhibits infinite-range statistical dependencies, despite being finite state. The Even Process consists of words in which 1s occur in blocks (pairs) of even length bounded by 0s. The evenness criterion is a statistical dependency of infinite range. One consequence is that the Even Process has infinite Markov order. Moreover, its irreducible forbidden set is countably infinite: $\mathcal{F} = \{01^{2n+1}0 : n \in \mathbb{Z}_{\geq 0}\}$.

There are two transient states A and B and the initial distribution is $\pi_0 = \{1, 0, 0, 0\}$. The state transition matrices for the generator are:

$$T^{(0)} = \begin{pmatrix} 0 & 0 & 1/3 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } T^{(1)} = \begin{pmatrix} 0 & 2/3 & 0 & 0 \\ 3/4 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The recurrent states C and D are visited with probabilities $\widehat{\pi} = \{0, 0, 2/3, 1/3\}$—which state distribution is time invariant after an infinite number of observations.

The informational analysis says the entropy rate is $h_\mu = 2/3\ \mathfrak{I}(1/2) = 2/3$ bits per emission. The statistical complexity is $C_\mu = \mathfrak{I}(2/3)$ bits, as is the excess entropy. These are all the same as for the Golden Mean Process. Nonetheless, it is clear that the Even Process, being infinite-Markov is qualitatively different from the Golden Mean Process. For related analyses see Ref. [50].

To start analyzing the Even Process' information processes, as above, we determine the time-dependent state distribution $p(t)$ explicitly:

$$\begin{aligned} p(t=0) &= \big(1, 0, 0, 0\big), \\ p(t=1) &= \begin{matrix} \big(\Pr(\mathcal{S}|x_{0:0} = 0) & = (0, 1/3, 0, 0) \ , \\ \Pr(\mathcal{S}|x_{0:0} = 1) & = (0, 0, 2/3, 0)\big) \end{matrix} \\ p(t=2) &= \begin{matrix} \big(\Pr(\mathcal{S}|x_{0:1} = 00) & = (0, 0, 1/6, 0) \ , \\ \Pr(\mathcal{S}|x_{0:1} = 01) & = (0, 0, 0, 1/6) \ , \\ \Pr(\mathcal{S}|x_{0:1} = 10) & = (0, 0, 1/6, 0) \ , \\ \Pr(\mathcal{S}|x_{0:1} = 11) & = (1/2, 0, 0, 0)\big) \end{matrix} \end{aligned}$$

$$\ldots \ .$$

This illustrates the transient phase—how the initial state probabilities begin to settle toward the asymptotic distribution $\widehat{\pi}$. This relaxation takes infinite time; reflecting the processes infinite Markov order. And this, in turn, means that the information processes—such as, $h_\mu(t)$ and $C_\mu(t)$ also take infinite-time to reach their asymptotic values.

Table VII presents an agent's informational and semantic analysis of the Even Process. It shows, in particular, that the degree of meaning continues to change during this relaxation. In fact, the information processes do not reach their asymptotic behaviors until the agent is synchronized. And, this only occurs once a 0 is observed—taking the state to C.

As above, we leave explicitly calculating the companion information processes as an exercise. Nonetheless, Fig. 12 presents several of its information processes: entropy rate $h_\mu(t)$, anticipated information rate ρ_μ, ephemeral rate r_μ, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

Observer's Analysis of Even Process				
State	Measurement x	Surprise $-\log_2 \Pr(x)$ [bits]	Semantic State σ: Meaning	Degree of Meaning $\Theta(x)$ [bits]
A	λ	Not Defined	No Measurement	0.0
A	1	0.585	B: Unsync	$0.585 \ldots \infty$
A	0	1.585	C: Sync	$1.585 \ldots 0.585$
B	1	0.415	A: Unsync	$0.585 \ldots \infty$
B	0	1.585	C: Sync	$1.585 \ldots 0.585$
C	1	1	D: Odd #1s	1.585
C	0	1	C: Even #1s	0.585
D	1	0	C: Even #1s	0.585
D	0	∞	A: Loose sync; Reset	0.0

TABLE VII. Information and Semantic Analysis of the Even Process. Recall Fig. 8(d). (Reproduced with permission from Ref. [59].)

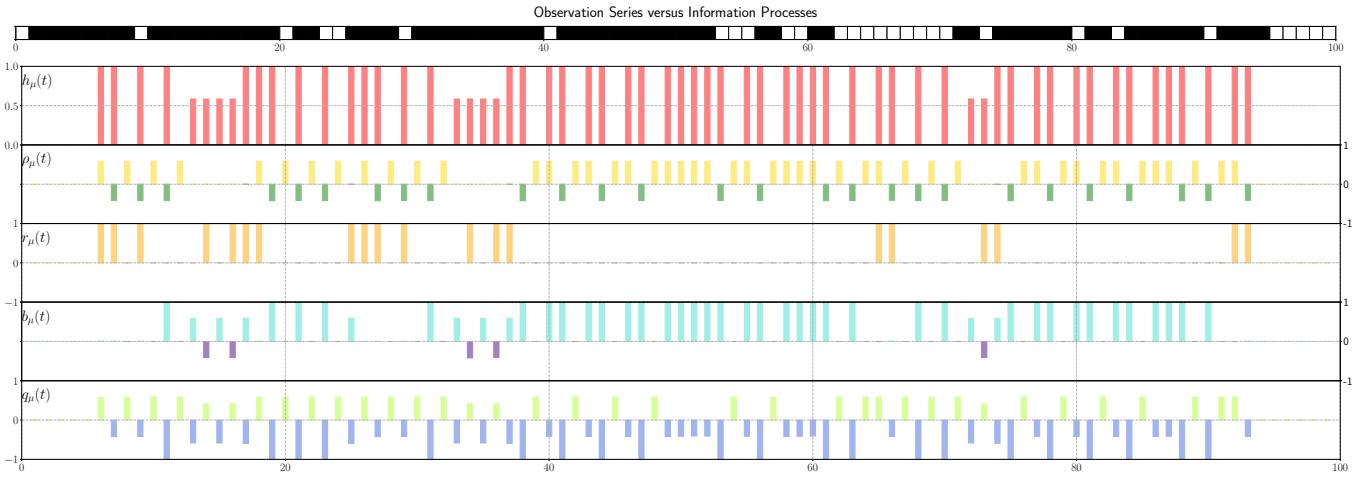

FIG. 12. Even Process observed time series sample versus several of its information processes: entropy rate $h_\mu(t)$, anticipated information rate $\rho_\mu(t)$, ephemeral rate $r_\mu(t)$, bound information rate $b_\mu(t)$, and enigmatic information rate $q_\mu(t)$.

In one sense, the Even Process is markedly more complex than the Golden Mean Process as it has infinite-range statistical properties. Specifically, while the Golden Mean as a single forbidden word $\mathcal{F} = 00$, the Even Process has a countable infinity $\mathcal{F} = \{0(11)^n 0 : n = 1, 2, \ldots\}$. In practical terms this enhances the complication of the informational narrative of the information processes given in Fig. 12.

In short, the Golden Mean's order-1 Markov order translates into the ability to track short words to determine the level of next-symbol unpredictability. Equivalently, in this case, tracking these short words allows one to know in which of the Golden Mean's causal states (B and C) the process is in at any given time.

This is not so for the Even Process. There are statistical dependencies of arbitrary length involved in knowing which state the process is. A quantitative consequence is the information measures at any given time t are not simply-interpreted values. However, we can easily extract interpretations of the Even Process' information process values—sometimes exactly, but typically only approximately. (That is, short of writing down these measures' closed form expressions using the methods of Ref. [58], which is the mandate for a sequel.)

Briefly, the easy interpretations are expedited by referring to the Even Process' state transition diagram in Fig. 8(d). First, we consider only asymptotic statistics by ignoring transient states A and B and concentrating on recurrent states C and D. Practically, one simply uses a realization far after the start state.

Notice that when in state C the next observation is a fair coin flip: $h_\mu(t) = 1$ bit. And, when in state D the next observation is wholly determined (it occurs certainly) to be a 1. It is the second 1 in the Even Process' characteristic pairs of 1s. That is, $h_\mu(t) = 0$ bit. Due to this $r_\mu(t)$ and $b_\mu(t)$ vanish, as seen in the information process plot. One can go much further, of course, but that is the burden of a sequel.

As with the Golden Mean Process, we again see the expected negative information measures. And, that the Even Process is a markedly more complex combination of the two information processing modes—Biased Coin and Periodic Processes-illustrated by the Golden Mean Process. Again, a sequel will provide a directly mecha-

Semantics of the Period-2 Process X as the Golden Mean Process X'				
State	Measurement x'	Δ Surprise $\log_2(\Pr(x)/\Pr(x'))$	State σ': Meaning	$\Delta\Theta = \log_2(\Pr(\sigma)/\Pr(\sigma'))$
A'	λ	0.0	No Measurement	0
A'	1	$-0.415 = \log_2((1/2)/(2/3))$	B': Sync	$-0.415 = \log_2((1/2)/(2/3))$
A'	0	$0.585 = \log_2((1/2)/(1/3))$	C': Sync	$0.585 = \log_2((1/2)/(1/3))$
B'	1	$-\infty = \log_2((0)/(1/2))$		Never Occurs
B'	0	$1 = \log_2((1)/(1/2))$	C': Isolated 0	1
C'	1	$0 = \log_2((1)/(1))$	B': Deterministic 1	0.585
C'	0	$-\infty = \log_2((0)/(1))$		Never Occurs

TABLE VIII. Semantics of Period-2 Process as if the Golden Mean Process. Recall Figs. 8(b) and (c).

Semantics of Even Process X as the Golden Mean Process X'				
State	Measurement x'	Δ Surprise $\log_2(\Pr(x)/\Pr(x'))$	State σ': Meaning	$\Delta\Theta = \log_2(\Pr(\sigma)/\Pr(\sigma'))$
A'	λ	Not Defined	No Measurement	0
A'	1	0	B': Sync	0
A'	0	0	C': Sync	0
B'	1	$0.585 = \log_2((3/4)/(1/2))$	B': Random 1	$-0.415 = \log_2((1/2)/(2/3))$
B'	0	$-1 = \log_2((1/4)/(1/2))$	C': Isolated 0	$0 = \log_2((1/3)/(1/3))$
C'	1	$-1 = \log_2((1/2)/(1))$	B': Deterministic 1	0.585
C'	0	$\infty = \log_2((1/2)/(0))$	A': Loose sync; Reset	0

TABLE IX. Semantics of Even Process as if the Golden Mean Process. Recall Figs. 8(c) and (d).

nistic explanation of how the information measures are being stored and transformed, making it clearer how the Even Process is more complex. Again, the proceeding's goal was an introduction to and initial exploration of information processes.

E. Misdirected Semantics

These example analyses all assumed the agent knew what the process generator was and that the agent came to be synchronized to the environment state. However, the semantic theory here applies more generally, to circumstances when the agent has an approximate or even incorrect internal model of its environment.

To be concrete, consider two examples of misinterpretation: An agent uses the Golden Mean ϵ-machine to interpret the semantics of (i) the Period-2 Process and (ii) the Even Process. (Recall Figs. 8(b), (c), and (d).) The analyses are presented in Tables VIII and IX.

We simply track how both processes, starting in their start state A, produce words up to length 2: $\omega \in \{00, 0, 110, 11\}$. The Biased Coin recognizes all; the Golden Mean disallows 00.

We use the information gain (or Kullback-Leibler divergence) to give the self-information difference between the expected next symbol versus actual and information distance between degree of meanings of interpreted state and actual state. The first relates to differing expectations in predicting symbols and transition probabilities—the *surprise* Δ:

$$\Delta = \log_2 \frac{\Pr(x)}{\Pr(x')} \ .$$

The second concerns the differing interpretations of the state semantics—$\Delta\Theta$:

$$\Delta\Theta = \log_2 \frac{\Pr(\rightarrow_x \sigma)}{\Pr(\rightarrow_{x'} \sigma')} \ .$$

Note that when comparing a process to itself these vanish: $\Delta = 0$ and $\Delta\Theta = 0$.

VIII. CONCLUSION

Many complex adaptive systems consist of an individual agent or a collection of interacting agents that take in and process data from their surroundings and then take actions based on what is gleaned. The preceding addressed only half of this, What kinds of stochastic process are involved as an agent monitors its environment? Our goal was to call out what is common—from an information-theoretic perspective—across such systems.

In this, we introduced information processes—time series of various kinds of Shannon information measure that capture a range of signal types from degrees of unpredictability to degrees of correlation and temporal memory. As a technical focus, we explored conditions for these signals' stationarity and ergodicity. These are properties that strongly affect what an agent can use statistically "downstream" to make reliable estimates of environmental

conditions and make robust decisions and take actions necessary for functioning.

Simply having well-behaved informational signals in hand, though, is far from the whole story of agent functioning. The question immediately arises as to what the signals mean to an agent and what the agent can do with them. We addressed the former by reviewing Shannon information measures, following on Refs. [50, 51]'s analysis of structured stochastic processes. This suite of measures has natural, informational, even functional meanings—they give an intrinsic semantics of an environment's behavior, as an agent experiences it. These statistics—environment unpredictability, the present's correlation with the past, memory required for optimal prediction, and the amount of hidden internal-state information—embody key kinds of information that an agent needs in order to properly operate.

Taken together then, information processes, their ergodicity, and their semantics lay a foundation for agentic information theory. That is, they comprise a first step to move beyond asymptotic properties to real-time signals needed for functioning, flourishing complex adaptive systems.

ACKNOWLEDGMENTS

The authors thank the Telluride Science Research Center for hospitality during visits and the participants of the Information Engines Workshops there. We thank Ryan James for his adept use of his Python *dit* discrete information theory package [62], Greg Wimsatt for comments on measure theory and stochastic processes, and Alex Boyd for the real-time thermodynamic data as a physical example of an information process. We thank them all for helpful discussions. This material is based upon work supported by, or in part by, the Art and Science Laboratory and the U.S. Army Research Laboratory and the U.S. Army Research Office under grant W911NF-21-1-0048.

Appendix A: Function of a Process is a Process

Proposition. Let $\{X(t)\}_{t\in\mathbb{Z}}$ be a stochastic process on probability space (Ω, F, P) and $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then $Y = \{Y(t) = f(X(t)), t \in \mathbb{Z}\}$, is a stochastic process.

Proof. First, $Y(t)$ is a random variable for each $t \in \mathbb{Z}$.

For any fixed $t \in \mathbb{Z}$, $Y(t) : \Omega \to \mathbb{R}$ is defined by $Y(t)(\omega) = f(X(t)(\omega))$. And, for any Borel set $B \subseteq \mathbb{R}$, its turns out that $Y^{-1}(t)(B) \in F$:

$$\begin{aligned} Y^{-1}(B) &= \{\omega \in \Omega : Y(t)(\omega) \in B\} \\ &= \{\omega \in \Omega : f(X(t)(\omega)) \in B\} \\ &= \{\omega \in \Omega : X(t)(\omega) \in B\} \\ &= X^{-1}(t)(f^{-1}(B)) \ . \end{aligned}$$

Since f is measurable, $f^{-1}(B)$ is a Borel set in $\mathbb{R}$.

Since $X(t)$ is a random variable—as $\{X(t) : t \in \mathbb{Z}\}$ is a stochastic process—then $X^{-1}(t)(f^{-1}(B)) \in F$.

Therefore, $Y^{-1}(t)(B) \in F$ for every Borel set B. That is, $Y(t)$ is a random variable.

Second, the collections of RVs forms a stochastic process.

Since $Y(t)$ is a RV for each $t \in \mathbb{Z}$, the collection $\{Y(t) : t \in \mathbb{Z}\}$ satisfies the definition of a stochastic process. And, its finite-dimensional distributions over subsets of indices and their probabilities are well-defined.

That is, the collection forms a stochastic process since f's measurability guarantees $f \circ X(t)$ preserves the RV property for each and all t. □

[1] M. Cross and H. Greenside. *Pattern Formation and Dynamics in Nonequilibrium Systems*. Cambridge University Press, Cambridge, United Kingdom, 2009.

[2] R. F. Streater. *Statistical Dynamics: A Stochastic Approach to Nonequilibrium Thermodynamics*. Imperial College Press, London, second edition, 2009.

[3] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75: 126001, 2012.

[4] J. Hertz, A. Krogh, and R. G. Palmer. *An Introduction to the Theory of Neural Networks*, volume 1 of *Lecture Notes, Studies in the Sciences of Complexity*. Addison-Wesley, Redwood City, California, 1991.

[5] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge, Cambridge, United Kingdom, 2003.

[6] H. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *Proc. 32nd Intl. Conf. Machine Learning*, 37:2256–2265, 2015.

[7] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. Statistical mechanics of deep learning. *Annu. Rev. Condens. Matter Phys.*, 11: 501–528, 2020.

[8] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[9] J. Rahme and R. P. Adams. A theoretical connection between statistical physics and reinforcement learning. *arXiv.org:1906.10228*, 2019. https://arxiv.org/abs/1906.10228.

[10] D. Melanson, M. A. Khater, M. Aifer M, K. Donatella, M. H. Gordon, T. Ahle, G. Crooks, A. J. Martinez, F. Sbahi, and P. J. Coles. Thermodynamic computing system for ai applications. *arXiv:2312.04836*, 2023.

[11] P. J. Coles, C. Szczepanski, D. Melanson, K. Donatella, A. J. Martinez, and F. Sbahi. Thermodynamic AI and the fluctuation frontier. *2023 IEEE Int. Conf. on Rebooting Computing (ICRC) (IEEE)*, pages 1–10, 2023.

[12] C. Bechinger, R. Di Leonardo, H. Lowen, R. Reichhardt, G. Volpe, and G. Volpe. Active particles in complex and crowded environments. *Rev. Mod, Phys.*, 88:004500, 2016.

[13] S. Sattari, U. S. Basak, R. G James, L. W. Perring, J. P. Crutchfield, and T. Komatsuzaki. Modes of information flow in collective cohesion. *Science Advances*, 8(6):1–13, 2022.

[14] A. Gupta, S. Savarese, S. Ganguli, and F. Li. Embodied intelligence via learning and evolution. *Nat. Commun.*, 12:5721, 2021.

[15] E. Bonabeau, G. Theraulaz, and M. Dorigo, editors. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, New York, 1999.

[16] H. A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, Massachusetts, third edition, 1996.

[17] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[18] C. G. Langton, editor. *Artificial Life*, Redwood City, California, 1989. Addison-Wesley.

[19] K. Maruyama, F. Nori, and V. Vedral. Colloquium: The physics of Maxwell's demon and information. *Rev. Mod. Phys.*, 81:1, 2009.

[20] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of Maxwell's Demon. *Proc. Natl. Acad. Sci. USA*, 109(29):11641–11645, 2012.

[21] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11:131–139, 2015.

[22] T. Sagawa. Thermodynamics of information processing in small systems. *Prog. Theo. Phys.*, 127(1):1–56, 2012.

[23] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New Journal of Physics*, 18:023049, 2016.

[24] A. M. Jurgens and J. P. Crutchfield. Taxonomy of prediction. *arXiv:2504.11371*, 2025.

[25] R. W. Yeung. *Information Theory and Network Coding*. Springer, New York, 2008.

[26] R. G. James, J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Many roads to synchrony: Natural time scales and their algorithms. *Physical Review E*, 89:042135, 2014.

[27] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[28] A. B. Boyd, J. P. Crutchfield, M. Gu, and F. C. Binder. Thermodynamic overfitting and generalization: Energetic limits on predictive complexity. *New Journal of Physics*, 27:063901, 2025.

[29] T. Vicsek, A. Czirok, W. Ben-Jacob, I. Cohen, and O. Shochet. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Let.*, 76(6):1226, 1995.

[30] E. Shaw. Schooling fishes. *American Scientist*, 66(2): 166–175, 1978.

[31] I. D. Couzin, J. Krause, N. R. Franks, and S. A. Levin. Effective leadership and decision-making in animal groups on the move. *Nature*, 433:513–516, 2005.

[32] L. Szilard. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.*, 53:840–856, 1929.

[33] D. Gier and J. P. Crutchfield. Intrinsic and measured information in separable quantum processes. *Entropy*, 27: 599, 2025.

[34] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017.

[35] J. P. Crutchfield and C. Aghamohammdi. Not all fluctuations are created equal: Spontaneous variations in thermodynamic function. *Entropy*, 26(11):894, 2024.

[36] V. S. Vijayaraghavan, R. G. James, and J. P. Crutchfield. Anatomy of a spin: The information-theoretic structure of classical spin systems. *Entropy*, 19:214, 2017.

[37] E. Schrodinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge, United Kingdom, 1944.

[38] S. E. Marzen and J. P. Crutchfield. Optimized bacteria are environmental prediction engines. *Physical Review E*, 98:012408, 2018.

[39] S. E. Marzen and J. P. Crutchfield. Prediction and dissipation in molecular sensors: Conditionally markovian channels driven by memoryful environments. *Bull. Math. Bio.*, 82(25):1–46, 2020.

[40] J. P. Crutchfield and A. M. Jurgens. Information machines—foundations. *in preparation*, 2025.

[41] S. Loomis and J. P. Crutchfield. Topology, convergence, and reconstruction of predictive states. *Physica D*, 445, 2021.

[42] P. Billingsley. *Probability and Measure*. Wiley Interscience, New York, third edition, 1995.

[43] R. M. Gray. *Probability, Random Processes, and Ergodic Theory*. Springer-Verlag, New York, second edition, 2009.

[44] W. Feller. *An Introduction to Probability Theory and its Applications*. Wiley, New York, third, revised edition, 1970.

[45] C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[46] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.

[47] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden Day, San Francisco, California, 1964.

[48] A. N. Kolmogorov. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR*, 124:754, 1959. (Russian) Math. Rev. vol. 21, no. 2035b.

[49] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[50] R. G. James, C. J. Ellison, and J. P. Crutchfield. Anatomy of a bit: Information in a time series observation. *CHAOS*, 21(3):037109, 2011.

[51] R. G. James, K. Burke, and J. P. Crutchfield. Chaos forgets and remembers: Measuring information creation, destruction, and storage. *Physics Letters A*, 378:2124–2127, 2014.

[52] R. G. James, J. R. Mahoney, and J. P. Crutchfield. Information trimming: Sufficient statistics, mutual information, and predictability from effective channel states. *Phys. Rev. E*, 95(6):060102, 2017.

[53] P. M. Ara, R. G. James, and J. P. Crutchfield. The elusive present: Hidden past and future dependence and why we build models. *Phys. Rev. E*, 93(2):022143, 2016.

[54] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, 2000.

[55] J. Sun and E. M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D*, 267:49–57, 2014.

[56] R. G. James, N. Barnett, and J. P. Crutchfield. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.*, 116(23):238701, 2016.

[57] N. Barnett and J. P. Crutchfield. Computational mechanics of input-output processes: Structured transformations and the ϵ-transducer. *J. Stat. Phys.*, 161(2):404–451, 2015.

[58] J. P. Crutchfield, P. Riechers, and C. J. Ellison. Exact complexity: Spectral decomposition of intrinsic computation. *Phys. Lett. A*, 380(9-10):998–1002, 2016.

[59] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317 – 359, Reading, Massachusetts, 1992. Addison-Wesley.

[60] C. C. Strelioff and J. P. Crutchfield. Bayesian structural inference for hidden processes. *Physical Review E*, 89: 042119, 2014.

[61] N. F. Travers. Exponential bounds for convergence of entropy rate approximations in hidden Markov models satisfying a path-mergeability condition. *Stochastic Proc. Appln.*, 124(12):4149–4170, 2014.

[62] R. G. James, C. J. Ellison, and J. P. Crutchfield. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018. doi: https://doi.org/10.21105/joss.00738.