Neuroscience of Consciousness, 2017, 3(1): nix019

doi: 10.1093/nc/nix019
Research article

An algorithmic information theory of consciousness

Giulio Ruffini*,†

Starlab Barcelona, Avda. Tibidabo 47bis, 08035 Barcelona, Spain and Neuroelectrics Corporation, 210 Broadway, Cambridge, MA 02139, USA

*Correspondence address. Starlab Barcelona, Avda. Tibidabo 47bis, 08035 Barcelona, Spain. Tel: +34 679 69 5777; Fax: +34 93 212 6445; E-mail: giulio.ruffini@starlab-int.com
†http://orcid.org/0000-0003-3084-0177

Abstract

Providing objective metrics of conscious state is of great interest across multiple research and clinical fields—from neurology to artificial intelligence. Here we approach this challenge by proposing plausible mechanisms for the phenomenon of structured experience. In earlier work, we argued that the experience we call reality is a mental construct derived from information compression. Here we show that algorithmic information theory provides a natural framework to study and quantify consciousness from neurophysiological or neuroimaging data, given the premise that the primary role of the brain is information processing. We take as an axiom that "there is consciousness" and focus on the requirements for structured experience: we hypothesize that the existence and use of compressive models by cognitive systems, e.g. in biological recurrent neural networks, enables and provides the structure to phenomenal experience. Self-awareness is seen to arise naturally (as part of a better model) in cognitive systems interacting bidirectionally with the external world. Furthermore, we argue that by running such models to track data, brains can give rise to apparently complex (entropic but hierarchically organized) data. We compare this theory, named KT for its basis on the mathematical theory of Kolmogorov complexity, to other information-centric theories of consciousness. We then describe methods to study the complexity of the brain's output streams or of brain state as correlates of conscious state: we review methods such as (i) probing the brain through its input streams (e.g. event-related potentials in oddball paradigms or mutual algorithmic information between world and brain), (ii) analyzing spontaneous brain state, (iii) perturbing the brain by non-invasive transcranial stimulation, and (iv) quantifying behavior (e.g. eye movements or body sway).

Key words: algorithmic information theory; Kolmogorov complexity; cellular automata; neural networks; complexity; presence; consciousness; structured experience; neural correlates of consciousness; PCI; LZW; tCS; tACS; TMS; EEG; MEG; fMRI; AI

Introduction

Characterizing consciousness is a profound scientific problem (Koch *et al.* 2016) with pressing clinical and practical implications. Examples include disorders of consciousness (Laureys 2005; Casali *et al.* 2013), locked-in syndrome (Chaudhary *et al.* 2017), conscious state *in utero* (Lagercrantz and Changeux 2010), in sleep and other states of consciousness, in non-human animals, and perhaps soon in exobiology or in machines (Koch and Tononi 2008; Reggia 2013). Here, we address the phenomenon of structured experience from an information-theoretic perspective.

Science strives to provide simple models that describe observable phenomena and produce testable predictions. In line with this, we offer here the elements of a theory of consciousness based on algorithmic information theory (AIT). AIT studies the relationship between computation, information, and (algorithmic) randomness (Hutter 2007), providing a definition for the information of individual objects (data strings) beyond statistics (Shannon entropy). We begin from a definition of cognition in the context of AIT and posit that brains strive to

Received: 19 August 2016; **Revised:** 24 July 2017. **Accepted:** 27 July 2017

model their input/output fluxes of information (I/Os) with simplicity as a fundamental driving principle (Ruffini 2007, 2009). Furthermore, we argue that brains, agents, and cognitive systems can be identified with special patterns embedded in mathematical structures enabling computation and compression.

A brief summary of what we may call the Kolmogorov theory of consciousness (KT) is as follows. We start from the subjective view ("my brain and my conscious experience"):

1. "There is information and I am conscious." Information here refers to the messages/signals traveling in and out of my brain or even within parts of my brain (I/O streams), and to Shannon's definition of the information conveyed by those messages.
2. "Reality, as it relates to experience and phenomenal structure, is a model my brain has built and continues to develop based on input–output information." The phenomenal structure of consciousness encompasses both sensory qualia and the spatial, temporal, and conceptual organization of our experience of the world and of ourselves as agents in it (Van Gulick 2016). Brains are model builders and compressors of information for survival. Cognition and phenomenal consciousness arise from modeling, compression, and data tracking using models. At this stage, from what really is a mathematical framework, "I, (my brain)" derive, from the available information and computation, concepts such as chair, mass, energy, or space (physics is itself a derived, emergent concept).
   Then we shift to the objective view: what kind of mathematical structures connecting the concept of information with experience could describe the above?
3. We argue that the proper framework is provided by AIT and the concept of algorithmic (Kolmogorov) complexity. AIT brings together information theory (Shannon) and computation theory (Turing) in a unified way and provides a foundation for a powerful probabilistic inference framework (Solomonoff). These three elements, together with Darwinian mechanisms, are crucial to our theory, which places information-driven modeling in agents at its core.
4. To make the discussion more concrete, we briefly discuss Cellular Automata (CA). These represent one example for the definition of information and computation, and of "brains" as special complex patterns that can actually represent (model) parts of the universe. CAs, as universal Turing machines (TMs), can instantiate embedded sub-TMs and provide an example of how complex-looking, entropic data chatter can be produced by simple, iterative rules. This provides a conceptual bridge to relate algorithmic complexity and other measures and "flavors" of complexity (e.g. entropy, power laws, fluctuation analysis, fractals, complex networks, and avalanches).
5. We return to the subjective and hypothesize that structured, graded, and multidimensional experience arises in agents that have access to simple models. These models are instantiated on computational substrates such as recurrent neural networks (RNNs) and are presumably found by successful agents through interaction with a complex-looking world governed by simple rules.
   Finally, based on the prior items and shifting to empirical application,
6. We examine methods to characterize conscious systems from available data (internal/physiological or external/behavior) and propose lines for further research.

We do not address here the "hard problem" of consciousness—the fundamental origin of experience (Chalmers 1995). We assume that "there is consciousness," which, with the right conditions, gives rise to structured experience, much as we assume that "there is a quantum electromagnetic field" with particular states we call photons. We focus instead on understanding how structured experience is shaped by the algorithmic characteristics of the models brains (or other systems) build with simplicity as a guiding principle. We aim to link the properties of models with those of experience, such as uniqueness, unity, and strength. In this sense, we are aligned with the idea that phenomenal structure requires complex representations of the world (as in representational theories of consciousness) (Van Gulick 2016), and also that we should address the "real problem" (Seth 2016): "how to account for the various properties of consciousness in terms of biological mechanisms; without pretending it doesn't exist (easy problem) and without worrying too much about explaining its existence in the first place (hard problem)." An important new element is that we study "mathematical" mechanisms that, as such, can potentially be generalized beyond biology. This is an ambitious but challenging program. In the Discussion section, we discuss some limitations and open questions.

Computation, Compression, and Cognition

The definition of a Universal TM (Turing 1936) provides the mathematical foundation for computation and information theory and hence plays a key role in KT. Although our starting point is mathematical, it is readily linked to physics. In practice, all formulations of fundamental physics theories can be set on mathematical frameworks in which there is a description of the universe called the "state" (a string) and dynamic laws (effective procedures) that transform the state in time (computation) through "recursion." The state can be fully described given sufficient information (it is literally a string)—both in classical and in quantum theories—and evolves computing its future (Lloyd 2002). The field of physics is guided by the notion that some simple laws dictate this evolution. A possible conclusion is summarized by the conjecture (called "digital physics") that the universe is discrete and isomorphic to a TM. Although the specific choice of a physical theory is not of immediate concern for us, KT is certainly aligned with the idea that the universe is isomorphic to—or can be fully described by—such a mathematical structure, and that organisms are examples of special complex patterns embedded in it with the interesting property of being capable of modeling parts of the universe. The statement that the universe is a TM is important, among other reasons, because TMs can represent/embed others—and KT adopts the notion that brains are such embedded sub-TMs in the universe. Both CAs and RNNs provide examples of TMs, which may be appropriate at different levels of description.

CAs are mathematical structures defined on a cell grid with simple local interaction rules (Wolfram 2002), and they encapsulate many of fundamental aspects of physics (spatiotemporal homogeneity, locality, and recursion). They can be used to formalize the concepts of computation, information, and emergence of complex patterns and have attracted a great deal of interest because they capture two basic aspects of many natural systems: (i) they evolve according to local homogenous rules and (ii) they can exhibit rich behavior even with very simple rules. The simplest interesting example is provided by a 1D lattice of binary-valued "cells," with nearest neighbor interaction. A rule specifies, for the next iteration (dynamics) the value at that location from its prior value and that of its neighbors (state). Surprisingly, some of these rules have been shown to produce universal computers—such as Rule 110 (Cook 2004). That is, the patterns such a simple system

generates can be used to emulate a universal TM (as is Conway's 2D Game of Life, Gardner 1970). The initial configuration of the CA provides the program (Wolfram 2002). CAs can produce highly entropic data, with power law behavior (Kayama 2010; Mainzer and Chua 2012; Ninagawa 2013). Thus, CAs or similar systems represent interesting frameworks to study measurable hallmarks of computation and compression, and establish links with other complexity "flavors" (as discussed, e.g. in Mainzer and Chua 2012). Although we will not attempt to do so here, we note that CAs may provide a mathematical framework to formalize the definition of information and interaction, as required in definition of "agent" below.

Neural networks (NNs) represent another important paradigm of computation with a direct application in cognitive neuroscience and machine learning. Feedforward networks have been shown to be able to approximate any reasonable function (Cybenko 1989; Hornik 1991). Remarkably, if the function to be approximated is compositional (recursive), then a hierarchical, feedforward network requires less training data than one with a shallow architecture to achieve similar performance (Mhaskar *et al.* 2016). Significantly, RNNs are known to be Turing complete (Siegelmann and Sontag 1995). Recurrence in NNs thus enables universal modeling. There is increasing evidence that the brain implements such deep, recursive, hierarchical networks—see, e.g. Taylor *et al.* (2015).

Cognition from information

In this section, we attempt to formalize our ideas. If all that brains have access to is information, we can naturally think of brains as "information processing machines"—computers in the mathematical sense (TMs)—and questions about our experience of reality should be considered within the context of AIT. Our "Input/Output streams (I/Os)" include information collected from visual, auditory, proprioceptive and other sensory systems, and outputs in the form of PNS mediated information streams to generate actions affecting the body (e.g. via the autonomic system) or the external world (e.g. body movements or speech). We will use the term "cognition" here to refer to the process of model building and model-driven interaction with the external world (Ruffini 2007). Since it is a crucial concept in KT, let us define more formally the notion of "model" (Fig. 1a):

Definition 1. A model of a dataset is a program that generates (or, equivalently, compresses) the dataset efficiently, i.e. succinctly.

As discussed in Ruffini (2016), this definition of model is equivalent to that of a classifier or generating function—NNs and other classifiers can be seen to essentially instantiate models. A succinct model can be used to literally compress information by comparing data and model outputs and then compress the (random) difference or error using, e.g. Huffman or Lempel–Ziv–Welch (LZW) coding (Kaspar and Schuster 1987; Cover and Thomas 2006). Also, a good model must be capable of accounting for a large repertoire of potential I/Os. For example, Newtonian physics is a simple model that accounts for kinematics, dynamics, and gravitational phenomena on the Earth (falling apples) and space (orbit of the Moon). Naturally, a powerful model is both comprehensive and integrative, encompassing multiple data streams (e.g. auditory, proprioceptive, and visual data). Examples of models built by brains include our concepts of space and time, hand, charge, mass, energy, coffee cups, quarks, tigers, and people.

To survive—to maintain homeostasis and reproduce—brains build models to function effectively, storing knowledge economically (saving resources such as memory or time). They use models to build other models, for agile recall and decision making, to predict future information streams, and to interact successfully with the world. Having access to a good, integrated model of reality with compressive, operative, and predictive power is clearly an advantage for an organism subjected to the forces of natural selection (from this viewpoint, brains and DNA are similar compressing systems acting at different time scales). Furthermore, when a brain interacts actively with the rest of the universe, it disturbs it with measurements or other actions (represented as information output streams). The information gathered from its inputs (senses) depends on how it chooses to extract it from the outside world (through the passive and active aspects of sensing or other actions). A more complete and therefore more useful model of reality of an active brain must include a model of itself—of "bodies" and internal "algorithms," for example. This creates a "strange loop" (Hofstadter 2007; Ruffini 2007) in terms of model representations. Such self-models correspond here to what are called body representation and self-awareness.

On the basis of the notion of modeling, we now define a cognitive system or "agent," of which a brain is an example:

Definition 2. A cognitive system or agent is a model-building semi-isolated computational system controlling some of its couplings/information interfaces with the rest of the universe and driven by an internal optimization function.

Figure 1b displays schematically the modeling engine and the resulting error stream from comparison of data and model outputs. These are passed onto an action module that makes decisions guided by an optimization function (possibly querying the model for action simulations) and generates output streams, which also feedback to the model. A classical thermostat or a machine-learning classifier are not agents by this definition, but new artificial intelligence systems being developed are. As an example, we refer to Bongard *et al.* (2006), where a four-legged robot uses actuation–sensation relationships to model its own physical structure, which it then uses to generate locomotion, or to the recent Deep Reinforcement Learning results, where deep learning and reinforcement learning are combined very much as in the figure to create AI systems that excel in Atari video-game universes (Mnih *et al.* 2015).

Simplicity and Kolmogorov complexity ($\mathcal{K}$)

Compression (and therefore simplicity) was formalized by the mathematical concept of algorithmic complexity or "Kolmogorov complexity" ($\mathcal{K}$) and co-discovered during the second half of the 20th century by Solomonoff, Kolmogorov, and Chaitin. We recall its definition: "the Kolmogorov complexity of a string is the length of the shortest program capable of generating it." More precisely, let $\mathcal{U}$ be a universal computer (a TM), and let p be a program. Then the Kolmogorov or algorithmic complexity of a string x with respect to $\mathcal{U}$ is given by $\mathcal{K}_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p)$, i.e. the length $l(p)$ of the shortest program that prints the string x and then halts (see e.g. Cover and Thomas 2006; Li and Vitanyi 2008). Crucially, although the precise length of this program depends on the programming language used, it does so only up to a string-independent constant. An associated useful notion is the "mutual algorithmic information (MAI)" between two strings (Grunwald and Vitanyi 2004), the algorithmic analog of Shannon mutual information.

We also need to point out a derived elegant concept, the "Kolmogorov Structure Function" of a dataset (Cover and Thomas 2006; Ruffini 2016), as well as the related concept of Effective Complexity (Gell-Mann and Lloyd 2003). Briefly, one

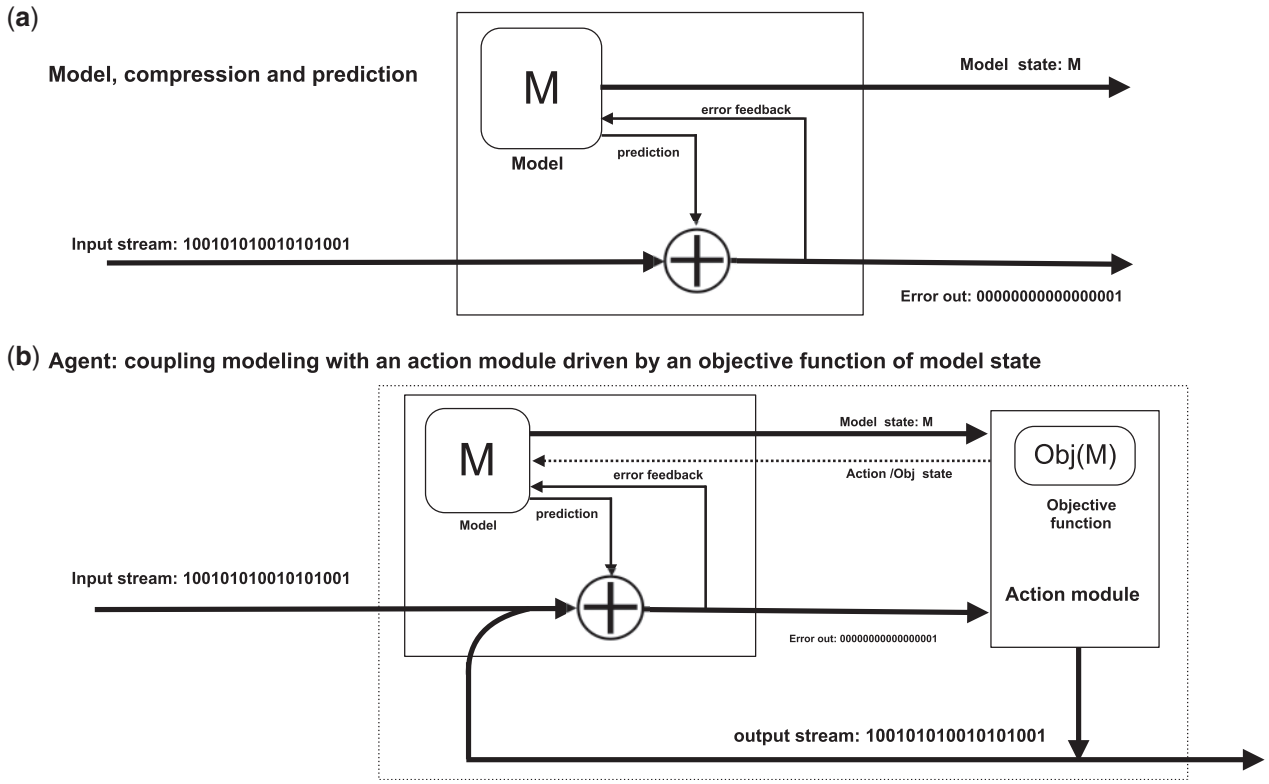

Figure 1. Top (a): Modeling for predictive compression. The arrows indicate information flow and the circled plus sign comparison/difference (XOR gate). Bottom (b): An agent with coupled modeling and an action modules. The action module contains an optimization objective function (e.g. homeostasis) and may include, e.g. a motor system, and will query the model (this time for "imagery") to plan and select the next actions. The model itself may be informed of the state of the action module directly (dotted line) or indirectly via the output stream (concatenated to the input stream). Without loss of generality a single stream is shown, although I/O streams represent multidimensional data.

can conceptually split the Kolmogorov optimal program describing a data string into two parts: a set of bits describing its regularities and another which captures the rest (the part with no structure). The first term is the effective complexity, the minimal description of the regularities of the data. This concept brings to light the power of the notion of Kolmogorov complexity, as it provides, single handedly, the means to account for and separate regularities in data from noise.

Gödel's incompleteness theorem, or its equivalent, Turing's halting theorem, implies that we cannot compute in general $\mathcal{K}$ for an arbitrary string: it is impossible to test all possible algorithms smaller than the size of the string to compress, since we have no assurance that the TM will halt (Chaitin 1995). However, within a limited computation scheme (e.g. in terms of programming language resources or computation time), variants of algorithmic complexity can be calculated. An example of this is Lempel–Ziv–Welch compression (Ziv and Lempel 1978), a simple yet fast algorithm that exploits the repetition of symbol sequences (one possible form of regularity). LZW file length is actually equivalent to entropy rate, an extension of the concept of entropy for stochastic sequences of symbols. LZW provides a useful if limited "upper bound" to $\mathcal{K}$.

The connection between simplicity, statistics, and prediction was developed by Solomonoff through the definition of the "algorithmic or universal probability $P_U(x)$ of a string x" (Li and Vitanyi 2008). This is the (prior) probability that a given string x could be generated by a random program. An important result is that $P_U(x) \approx 2^{-\mathcal{K}_U(x)}$. Thus, the probability of a given string being produced by a random program is dominated by its Kolmogorov complexity. Because of this, a Bayesian prior for simplicity may be a good strategy for prediction, e.g. in a universe where data are generated by ferrets typing programs or by other random program generating mechanisms. Although we can only hypothesize the existence of such a data generating process, we do seem to inhabit a universe described by simple rules. Thus, we will assume here that while I/O streams encountered by agents may appear to be complex (entropic), they are inherently simple, allowing for compression (the deterministic, "simple physics hypothesis"). From this, and from considerations on the evolutionary pressure on replicating agents (natural selection favoring pattern-finding agents), we formulate the following hypothesis:

Hypothesis 1. Successful replicating agents find and use simple models of their I/Os.

As far as an agent can tell, "reality" is the simplest program it can find to model data-streams generated from its interaction with the world.

Consciousness and KT

We address next the nature of conscious content. In what follows, we assume that there is a strong link between structured experience and cognition, the cognitive substrates, and processes involved in modeling I/Os.

Structured consciousness requires compressive models of I/Os

From a cognitive perspective, we have argued that what we call reality is represented and shaped by the simplest programs

brains can find to model their interaction with the world. In some sense, simplicity is equivalent to reality and therefore, we now hypothesize, to structured experience. When we become conscious of something, we become conscious of it through a model, which is chosen among many as the one best fitting the available data. In more detail, we propose our next hypothesis, relating cognition and consciousness:

Hypothesis 2. Structured conscious content is experienced by agents tracking I/Os using successful, simple models. The more compressive these models are, the stronger the subjective structured experiences generated.

In other words, conscious experience has a richer structure in agents that are better at identifying regularities in their I/Os streams, i.e. discovering and using more compressive models. In particular, a "more" conscious brain is one using and refining succinct models of coherent I/Os (e.g. auditory and visual streams originating from a common, coherent source, or data accounting for the combination of sensorimotor streams). We may refer to this compressive performance level as "conscious level." It is ultimately limited by the algorithmic complexity of the universe the agent is in and the resources it has access to.

Returning to Fig. 1, the better the fit of the model with all available data (integrating present and past multisensory streams), the stronger the experience (how real it will feel to the agent) and the stronger the impact on behavior. The model itself is a mathematical, multidimensional, highly structured object, and can easily account for a huge variety of experiences. It will also, in general, be compositional and recursive (assuming those are properties of I/Os). An implicit element here is thus that consciousness is a unified, graded, and multidimensional phenomenon.

Let us clarify that here highly compressive implies "comprehensive," i.e. that all the I/O data streams available up to the moment of experience are ideally to be accounted for, and that "compressive" refers to the length of model plus (compressed) error stream being short. Past I/Os (possibly encoded in the form of prior models), play an important role: the algorithmic complexity of new data given available old data must be low (simple).

In KT, structured conscious awareness is thus associated to information processing systems that are efficient in describing and interacting with the external world (information). An ant, e.g. represents such a system. Furthermore, some experiences may require a self-awareness, as we discussed before: if the appropriate model has to take into account the agent's actions (the output streams), then self-awareness (self-modeling) will become an important element of structured experience. However, not all interactions may call for a self-model (e.g. passively perceiving an object may not require running a self-model, while dancing presumably does). Self-modeling includes here all the agent's elements (e.g. including the Action module policy in the figure).

In Ruffini (2009), we hypothesized a related conjecture with regard to the experience of "Presence," the subjective experience of being somewhere. We may view this phenomenon as a consequence of our prior hypotheses: "Given a set of models for available data, an agent will select the most compressive one, or equivalently, the model that will feel most real." Again, by data here we mean all available data up to the present moment, some of which may be from, e.g. the past hour, or encoded in models built from much older interactions.

Apparent complexity from simplicity

Can we associate the characteristics of electrophysiological or metabolic spatiotemporal patterns in brains to conscious level? Although somewhat counterintuitive, in KT agents that run simple models in conscious brains may appear to generate (Shannon) apparently complex data. By "apparently complex data" streams, we mean those that are inherently simple yet entropic and probably hard to compress by weak algorithmic complexity estimators such as LZW. The context for this apparent paradox is the aforementioned hypothesis (the deterministic, simple physics hypothesis) that the universe is ruled by simple, highly recursive programs which generate entropic data. As Mandelbrot, Wolfram, and others have shown, apparently complex data streams can be generated by very simple, recursive special models (Wolfram 2002) (called "deep" models in Ruffini 2016). By this we mean models such as an algorithm for the digits of π, which are not compressed by algorithmic complexity estimators such as LZW. In such a world, a brain tracking—and essentially simulating—high entropy data from its interaction with the world will itself produce complex looking data streams.

Recapping, we hypothesize that driven by natural selection in a complex looking but intrinsically simple universe, replicating agents running and developing models of reality will instantiate recursive computation (that being necessary for deep modeling), running compressive, "deep" programs. The data produced by such recursive agents can display features of critical systems ("order at the edge of chaos") situated between the kingdoms of simple and random systems (Li and Nordahl 1992). Simple, deep programs will model and therefore generate entropic, fractal-looking data, and one whose structure is characterized by power laws, small world (Gallos *et al.* 2012) or scale-free networks (Eguiluz *et al.* 2005) associated with the hierarchies in the systems we find in the natural world (West 1999; Albert and Barabasi 2002; Ravasz and Barabasi 2003; He 2014). While a brain capable of universal computation may produce many different types of patterns—both simple (e.g. constant or repetitive) and entropic—a healthy brain engaging in modeling and prediction of complex I/Os will produce complex-looking, highly entropic data. Such "apparent complexity" is what is evaluated by entropy or LZW compression measures of, e.g. electrophysiological or metabolic brain data (Casali *et al.* 2013; Schartner *et al.* 2015; Andrillon *et al.* 2016; Hudetz *et al.* 2016; Schartner *et al.* 2017). First-order entropy, entropy rate, or LZW provide "upper bounds" to the algorithmic complexity of such data. We summarize this as follows:

Consequence 1. Conscious brains generate apparently complex (entropic) but compressible data streams (data of low algorithmic complexity).

Thus, in principle, the level of consciousness can be estimated from data generated by brains, by comparing its apparent and algorithmic complexities. Sequences with high apparent but low algorithmic complexity are extremely infrequent, and we may call them "rare sequences." Healthy, conscious brains should produce such data. Although providing improved bounds on algorithmic complexity remains a challenge, an apparently complex data stream generated from a low algorithmic complexity model should in principle be distinguishable from a truly random one, leaving traces on metrics such as entropy rate, LZW, power law exponents and fractal dimension. If brain data are generated by a model we know (e.g. one fixed in an experimental scenario), a better bound for its algorithmic complexity could be derived by showing that the model can be used to further compress it. As an example, consider a subject whom we ask to imagine, with eyes closed, parabolic trajectories from cannonballs. Using Newton's equations, we should be able to compress the subject's EEG data beyond

LZW, demonstrating that it is partly generated by an internal physical model. As discussed, such apparent complexity from simplicity points to the EEG data being generated by deep programs embedded in biological networks that are modeling real work data.

MAI between world and agent

A related consequence of the above is that the MAI between world and brain generated data should be high. A model is a compressed representation of the external world. The actual program length instantiated in the agent should be much shorter than raw world data, while the MAI between both program/model and data should be high. We may also expect, in addition, that world data will not be obviously simple (i.e. entropic and not yet in compressed form). A simple example is the use of electrophysiology or fMRI data to reconstruct images projected on the retina (Stanley *et al.* 1999; Nishimoto *et al.* 2011). A more interesting case is when there exists at least one neuron that fires exclusively when an instance of a percept is presented (corresponding to a good model being run), such as in "grandmother" cells (Quiroga *et al.* 2005). Indeed, the information stemming from such a cell would allow us to compress the input stream more effectively.

Consequence 2. Consider a compressible ($l(x)/\mathcal{K}(x) > 1$) input data stream x and agent response data y as measured by, e.g. neuroimaging or agent behavior. In a conscious agent processing x (attending to it) the MAI $I_{\mathcal{K}}(x:y)$ will be high. Furthermore, the information about x in y will be in compressed form.

Note that a high MAI is a "necessary," not sufficient condition. High MAI between an external visual input and the state of the optical nerve or thalamus is also expected in a subject with eyes open. Our hypothesis is that information will be compressed in the cortex and present even if sensory inputs are disconnected, represented as a model—e.g. run as the subject imagines a visual scene. As models are presumably implemented in synaptic connectivity and neuronal dynamics, compressed representations of past input streams will be present in neuroimaging data. It is in this sense that we expect MAI between world and agent to increase with its actual or potential conscious level.

Relation to integration information, global workspace, and predictive processing theories of consciousness

KT is closely related to theories of consciousness that place information at their core, and it actually provides conceptual links among them. In Integration Information theory (IIT), the most important property of consciousness is that it is "extraordinarily informative" (Tononi and Koch 2008). It maintains that when we experience a conscious state, we rule out a huge number of possibilities. KT is strongly related to but not equivalent to ITT. KT places the emphasis on the use of simple models to track I/Os, which lead to structured experience and which we may call the mathematical substrates of consciousness. IIT emphasizes causal structure of information processing systems and the physical substrate of consciousness. However, the concept of a simple "model" (as defined above) may provide a more fundamental—or alternative—origin of the notion of a causal "complex" (a strongly interlinked causal information structure, Tononi *et al.* 2016). KT agrees well with other aspects of IIT. If structured experience is shaped by models, our belief in a particular model (as driven by the I/O streams up to this moment) efficiently rules out—or lowers our belief in—all other models for the experienced information streams. IIT emphasizes that information associated to a conscious state must be "integrated": the conscious state is an integrated whole that cannot be divided into sub-experiences (data from the I/Os must be tightly bound together). KT provides a mechanism for binding of information: a good, succinct model will by definition integrate available information streams into a coherent whole. While IIT states that "the level of consciousness of a physical system is related to the repertoire of causal states (information) available to the system as a whole (integration)," KT would say that the potential level of consciousness of a physical system is dictated by its ability to model its I/Os in an efficient manner. Economy of description implies both a vast repertoire (reduction of uncertainty or information) and integration of information. We note that simple programs (in the limit of Kolmogorov) are irreducible and Platonic mathematical objects (as in, e.g. "a circle is the set of points equidistant from another point"). This is another link with IIT and its central claim that an experience is identical to a conceptual structure that is maximally irreducible intrinsically.

We can establish closer links between KT and IIT by focusing on efficient NNs for the instantiation of models. By definition, the model encoded by a network specifies which value holders (nodes) to use, how to connect them, and an initial state. The result may be "integrated" or not. Loosely, if it is an effective (simple) encoding, we would expect interconnectivity and intercausality in the elements of the network. It turns out that we should also expect that perturbations of nodes of such a network, when activated in detecting a matching pattern (i.e. running a model), will propagate further in the network, as found in Casali *et al.* (2013) (Ruffini 2016).

Global workspace theory (GWT) (Baars 1988; Dehaene *et al.* 2003) has common elements with IIT and KT. It states that "conscious content provides the nervous system coherent, global information" (Baars 1983), i.e. what we call in KT a (global) model. KT and IIT are in some sense meta-theories, with the biological brain (and thus perhaps GWT) as a particular case. The fact that effective models may require parallel information processing in KT maps into GWT's requirement that many areas of the brain be involved in those conscious moments in GWT. Since the original work of Baars, Dehaene and others (see, e.g. the recent results in Godwin *et al.* 2016) have identified global brain events to correspond to the conscious experience in numerous experiments. According to KT, the experience is associated to successful modeling validation events. Crucially, such events require integration of information from a variety of sensory and effective systems that must come together for model validation. Data must thus flow from a variety of sub-systems involving separate brain areas and merge—perhaps at different stages—for integrative error checking against a running model. There may be several such integrative sub-nodes (as in "grandmother" cells), whose outputs may themselves be integrated in a "grand model node." A candidate for such a location is the temporo–parietal–occipital junction (TPJ), an association area that integrates information from auditory, visual, and somatosensory information, as well as from the thalamus and limbic system (Koch *et al.* 2016).

Predictive processing theory (PP) (Friston 2009; Clark 2013; Hohwy 2013; Seth 2013, 2014) is also closely related to KT, with a focus on the predictive efficiency afforded by simple models of I/Os. It maintains that to support adaptation, the (Bayesian) brain must discover information about the likely external causes of sensory signals using only information in the flux of the sensory signals themselves. According to PP, perception solves this problem via probabilistic, knowledge-driven inference on the causes of sensory signals. In KT, the causes are formally

defined by models that are derived from the objective of compressing I/Os (and which include Bayesian modeling as a byproduct) in a computational framework, providing links with recursion and complexity measures.

Experimental Methods in KT

Modulating algorithmic complexity of input streams

In the case of the experience of Presence, consistency in the I/Os, in the sense of there being a simplifying low-level model available to match sensorimotor data, is a crucial element to enhance this experience ("place illusion," see Ruffini 2009; Slater 2009). As we progress higher in the modeling hierarchy, Bayesian prior expectations play an important role: explanations with a better match with past models are inherently simpler (leading to "Plausibility"). Virtual reality (VR) technology offers a powerful way to induce and manipulate Presence. KT (Hypothesis 1) predicts that given available models for existing data (past and present), the simplest will be chosen (Ruffini 2009).

Binocular rivalry is a well-established paradigm in the study of the neuroscience of consciousness. Briefly, two different images are presented to each eye and the experience of subjects typically fluctuates between two models, i.e. seeing the right or left image (Blake and Logothetis 2002; Blake and Tong 2008). According to Hypothesis 1, given this data stream and past ones, the subject's brain will select the simplest model it can find, which will then be experienced. First, we note that this experimental scenario breaks the subject's past models of the geometry of 3D space. Any given model of an object in 3D space will only match part of data stream (e.g. from a single eye). Since the subject does not have access to a simple model from past experience that integrates both retinal inputs, a partial model will be selected if the images are equally simple: the subject will use a model of one of the images and become conscious of only that particular image (the dominant one), discarding the other retinal inputs from conscious access (but may also patch both images up as in Kovács *et al.* 1996). KT suggests further dominance experiments in which the two images differ in terms of their simplicity, some of which have already been carried out. For example, natural images (with amplitude spectra of the form $A(f) \sim 1/f$) dominate (Baker and Graf 2009)—in KT because they agree better with available prior models, or, e.g. recognizable figures dominate over patterns with similar psychophysical traits, while upside down figures dominate relatively less (Yu and Blake 1992). Stimulus strength (luminance, contrast, motion Blake and Logothetis 2002) also play a role in KT, because strength typically relates to higher signal to noise ratio, which makes data streams more compressible than others. We can consider images that differ in their visual algorithmic complexity, e.g. the image of a regular versus an irregular polygon or target/context consistent (simpler) images, which dominate over inconsistent ones (Fukuda and Blake 1992) or, e.g. in a setting where at some point during an immersive VR experience two different images are presented to the subject, one congruent with the ongoing experience (more plausible), and the other less fitting with the overall experience. Perhaps, the subject can touch one of the objects appearing in the images or hear sounds associated to it. The prediction is that the subjects will tend to see the congruent (simpler model) image more often. According to KT, image training (prior model building) also leads to dominance (e.g. Dieter and Tadin 2016).

A direct approach to test the hypothesis that consciousness level self-reports correlate with the capacity of rule-finding (Hypothesis 2) is to prepare sensorial inputs of varying algorithmic complexities and assess the response of the brain (EEG or MEG) to rule-breaking (deviant inputs). This is the so-called "oddball paradigm" (Näätänen *et al.* 1978; Grau *et al.* 1998). The appearance of a surprise response to rule breaking is directly related to pattern detection (compression). According to KT, the level of response to a deviant input is associated with the complexity of the sequence and the available modeling power of the brain. Oddball experiments using patterns with varying complexity (including multimodality) could thus shed light on the role of conscious level and compression. For the purposes of studying higher level, structured consciousness, it may be more appropriate to work with the later parts in the EEG event-related potential (ERP), i.e. the P300b (Bekinschtein *et al.* 2009). We would expect that complex patterns might elicit weaker or delayed responses (in agreement with Atienza *et al.* 2003) and other experiments such as Kanoh *et al.* (2004) or take longer to learn (e.g. Benidixen and Schröger 2008 discuss how rapidly simple but abstract rules are learned). This is also addressed in Bekinschtein *et al.* (2009) and Faugeras *et al.* (2011, 2012) using the so-called "local-global" paradigm, although the authors' interpretation of the results (experience of global rule breaking requires awareness of stimuli) refers to the affordance of sustained perceptual representation. In KT, the interpretation is that the global rules used are algorithmically more complex than the local ones. Working memory is necessary for modeling, but not sufficient. This methodology has now been extended to the macaque brain, highlighting the role of a frontoparietal network in global regularity violation (Uhrig *et al.* 2014), including the activation of the temporoparietal area. KT would therefore predict that global-rule violation detection should not be available in situations in which subjects do not report experience (deep sleep, unresponsive wakefulness state, anesthesia, etc.). Furthermore, it suggests the exploration of stimulation sequences of increasing algorithmic complexity.

Perturbing cortical networks using brain stimulation

Massimini *et al.* (2005) used transcranial magnetic stimulation (TMS) to characterize changes of functional cortical connectivity during sleep. Later, Casali *et al.* (2013) used TMS similarly to generate propagating action potentials, with resulting EEG responses compressed using LZW to define a "perturbation complexity index (PCI)." The interpretation in IIT is that a high PCI reflects information (LZW) and integration (since the neural response originates from a common source and is therefore integrated by default). The interpretation in KT is slightly different, but in agreement with the idea that a PCI is indicative of conscious level. According to KT, brains run the simplest models they can find to track world data and make predictions. Such models, if implemented efficiently in NNs, should be quite sensitive to perturbations of their nodes while engaged in a task (Ruffini 2016)—disturbances should travel further, as they appear to do (Massimini *et al.* 2005). Moreover, we may expect that although EEG perturbations originate from a common cause (a localized TMS pulse), they will be represented differently across the cortex after non-linear propagation in cortical networks and will therefore be hard to compress using LZW, since LZW is quite limited in detecting and exploiting the potentially high MAI in the signals from different cortical sources for compression. We note that other, more powerful estimators of algorithmic complexity metrics can be explored. In addition, various non-invasive stimulation methods, such as transcranial current stimulation (tCS) can be used to generate sub-threshold stimulation-related potentials (SRPs) and study their complexity.

The complexity of spontaneous brain state

Suppose we collect multichannel spontaneous EEG data from a subject during a few seconds. An awake or sleeping brain during REM is characterized by fairly similar EEG: visually complex, fractal and distinct across channels and frequencies. The deeply asleep brain is dominated by slower rhythms with staccato-like bursts. The epileptic brain, the anesthetized brain, and the unresponsive brain all display less complex-looking EEG. We seek metrics that can differentiate between such data in terms of complexity and discriminate among healthy TM-like chatter from other forms of noise (Consequence 1). Using raw EEG, one may simply attempt to compress the data file from just a few seconds, e.g. using LZW. This technique has been shown to be useful already in a handful of examples—e.g. during anesthesia (Schartner *et al.* 2015) or sleep (Andrillon *et al.* 2016). Furthermore, we can derive connectivity networks in electrode or in cortical space and estimate their algorithmic complexity (see, e.g. Ray *et al.* 2007; Soler-Toscano *et al.* 2014; Zenil *et al.* 2014; Zenil *et al.* 2015a). Power laws and scale-free behavior with $1/f^{\alpha}$ spectra are also probably closely associated with simple TM chatter (Eguiluz *et al.* 2005), as proposed above. It is also known that hierarchical modular architecture of a network (its structure, as in the cortex) can deliver hallmark signatures of dynamical criticality—including power laws—even if the underlying dynamical equations are simple and non-critical (Friedman and Landsberg 2013). Although certainly of interest, further work is needed to make clear statements about the relation of apparent complexity (as measured by, e.g. LZW) and conscious level. For example, a random number generator produces maximal entropy, but its mutual information with the world is null. Thus, high apparent complexity alone does not necessarily imply high conscious level—it is necessary but not sufficient. Although a healthy brain running simple models is expected to produce apparently complex physiological chatter, we may be able to compress it beyond LZW if we have access to its underlying model, e.g. by controlling the experimental scenario to have a subject "run" a simple model—thus deriving better bounds on its algorithmic complexity.

The complexity of brain outputs

With regard to Consequence 1, behavior (an agent's output, such as hand-reaching motion, voice, gait, or eye movements) can be quantified in terms of apparent complexity. For example, Manor *et al.* (2010) studied postural sway (center-of-pressure dynamics) during quiet standing and cognitive dual tasking, and derived a complexity index using multiscale entropy (MSE) (Costa *et al.* 2002). MSE has also been used to classify human and robot behavior on Twitter, as in He *et al.* (2014). REM vs. NREM sleep eye movements provide another example. Eye movements have also been studied using entropy metrics, e.g. in autism spectrum disorder (Shic *et al.* 2008; Pusiol *et al.* 2016). More generally, the MAI between sensory inputs, brain state, and then behavioral response (I/Os) should correlate with consciousness level and awareness of the world (Consequence 2).

Discussion

KT proposes a mathematical framework to study cognition and consciousness based on AIT, where the conceptual kernel is the Kolmogorov complexity of a string. AIT provides the tools to study computation and compression of data from an apparently complex but intrinsically simple world. It takes as an axiom that "there is consciousness" and provides requirements for structured experience: it is only possible in computational systems such as brains that are capable of forming compressed representations of the world. The availability of compressive models gives rise to structured conscious phenomena in a graded fashion. Self-awareness is seen to arise naturally as a better model in agents interacting bidirectionally with the external world. We have thus linked by definition "conscious level" to the ability of building and running compressive models that generate "structured experience" (Hypothesis 2). While this can be seen as a limitation, in our view it provides a quantitative approach for the study of such elusive concepts.

KT holds that apparent complexity with hidden simplicity is the hallmark of data generated by agents running models of reality—cognitive systems enjoying structured experience, because the world is complex only in appearance. This provides a link between the conscious properties of systems and observables (e.g. EEG, fMRI time series, or behavior). We have argued that since brains track (or imagine) and model the external world (producing structured experience as a result), the apparent complexity (as measured by, e.g. entropy or LZW) but inherent simplicity of brain data (as measured by yet to be developed improved bounds on K) as well as the MAI of world data with present or past external brain inputs and outputs constitute key elements or the development of metrics of consciousness. However, more precise statements should be possible: the connections between algorithmic complexity and recursion with other complexity measures (e.g. power-laws, small-world network behavior, and fractal dimensions) remain to be fully established. CAs and RNNs may be good models to study these links. In addition to such research in mathematics, fundamental research in machine learning (e.g. studying the role of composition and simplicity in NNs) and in physics (studying how simple recursive laws lead to simple, recursive and deep effective theories at larger, coarse-grained scales) is needed to create stronger ties between mathematics, physics, cognitive neuroscience and artificial intelligence. In KT, life, cognition, and consciousness are all closely interrelated, graded phenomena united by the common thread of computation and compression in a complex, competitive environment. Even if KT is only partly correct, AIT-derived metrics should exhibit discriminatory power for the classification of conscious states and, importantly, a starting point for the generalization of our understanding of cognition and consciousness beyond biology.

Supplementary data

Supplementary data is available at *Neuroscience of Consciousness Journal* online.

Acknowledgements

This work has greatly benefited from discussions with many people, including Carles Grau, Ed Rietman and Walter Van de Velde, and has been partly supported by the FET Open Luminous project (H2020-FETOPEN-2014-2015-RIA under agreement No. 686764) as part of the European Union's Horizon 2020 research and training program 2014–2018. There is no data associated to this paper.

Funding

This research has partly been undertaken under the umbrella of the European FET Open project Luminous. This project has received funding from the European Union's

Horizon 2020 research and innovation programme under grant agreement No 686764.

References

Albert R, Barabasi A-L. Statistical mechanics of complex networks. *Rev Mod Phys* 2002;**74**.

Andrillon T, Poulsen AT, Hansen LK, *et al*. Neural markers of responsiveness to the environment in human sleep. *J Neurosci* 2016;**36**:6583–96.

Atienza M, Cantero JL, Grau C, *et al*. Effects of temporal encoding on auditory object formation: a mismatch negativity study. *Cognit Brain Res* 2003;**16**:359–71.

Baars B, (1983). Conscious contents provide the nervous system with coherent, global information. In: Davidson, R, Schwartz, G and Shapiro, D (eds). *Consciousness & Self-regulation*. New York: Plenum Press.

Baars B, (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baker DH, Graf EW. Natural images dominate in binocular rivalry. *PNAS* 2009;**106**.

Bekinschtein TA, Dehaene S, Rohaut B, *et al*. Neural signatures of the conscious processing of auditory regularities. *PNAS* 2009; **106**:1672–7.

Benidixen A, Schröger E. Memory trace formation for abstract auditory features and its consequences in different attentional contexts. *Biol Psychol* 2008;**78**:231–41.

Blake R, Logothetis NK. Visual competition. *Nat Rev Neurosci* 2002;**3**:13–21.

Blake R, Tong F. Binocular rivalry. *Scholarpedia* 2008;**3**:1578.

Bongard J, Zykov V, Lipson H. Resilient machines through continuous self-modeling. *Science* 2006;**314**:1118.

Casali AG, Gosseries O, Rosanova M, *et al*. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 2013;**5**:1–14.

Chaitin GJ. Randomness in arithmetic and the decline and fall of reductionism in pure mathematics. In: Cornwell J (ed.), *Nature's Imagination*. Oxford (UK): Oxford University Press, 1995, 27–44.

Chalmers D. Facing up to the problem of consciousness. *J Conscious Stud* 1995;**2**:200–19.

Chaudhary U, Xia B, Silvoni S, *et al*. Brain–computer interface–based communication in the completely locked-in state. *PLOS Biol* 2017;**15**:1–25.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 2013;**36**:181–204.

Cook M. Universality in elementary cellular automata. *Complex Syst* 2004;**15**:1–40.

Costa M, Goldberger AL, Peng C-K. Multiscale entropy analysis of complex physiologic time series. *Phys Rev Lett* 2002;**89**:1–4.

Cover TM, Thomas JA. *Elements of Information Theory*, 2nd edn. Hoboken, New Jersey: John Wiley & sons Inc., 2006.

Cybenko G. Approximations by superpositions of sigmoidal functions. *Math Control Signals Syst* 1989;**2**:303–14.

Dehaene S, Sergent C, Changeux J. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *PNAS* 2003;**100**:8520–5.

Dieter KC, Tadin MDMD. Perceptual training profoundly alters binocular rivalry through both sensory and attentional enhancements. *PNAS* 2016;**113**:12874–79.

Eguiluz VM, Chialvo DR, Cecchi GACA, *et al*. Scale-free brain functional networks. *Phys Rev Lett*, 2005;**94**:1–4.

Faugeras F, Rohaut B, Weiss N, *et al*. Probing consciousness with event-related potentials in the vegetative state. *Neurology* 2011;**77**:264–8.

Faugeras F, Rohauta B, Weiss N, *et al*. Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. *Neuropsychologia* 2012;**50**:403–18.

Fekete T, van Leeuwen C, Edelman S. System, subsystem, hive: boundary problems in computational theories of consciousness. *Front Psychol* 2016;**7**:1–17.

Fischer DB, Boes AD, Demertzi A, *et al*. A human brain network derived from coma-causing brainstem lesions. *Neurology* 2016; **87**:2427–34.

Fredkin E. An introduction to digital philosophy. *Int J Theoret Phys* 2003;**42**:189–247.

Fredkin E. Five big questions with pretty simple answers. *IBM J Res Dev* 2004;**48**

Friedman EJ, Landsberg AS. Hierarchical networks, power laws, and neuronal avalanches. *Chaos* 2013;**23**:013135. (

Friston KJ. The free-energy principle: a rough guide to the brain? *Trends Cogn Sci* 2009;**13**:293–301.

Fukuda H, Blake R. Spatial interactions in binocular rivalry. *J Exp Psychol Hum Percept Perform* 1992;**18**:362–70.

Gallos LK, Makse HA, Sigman M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *PNAS* 2012;**109**:2825–30.

Gardner M. Mathematical games–the fantastic combinations of John Conway's new solitaire game "life". *Sci Am* 1970;**223**: 120–3.

Gell-Mann M, Lloyd S. Effective complexity. In: Gell-Mann M and Tsallis C (eds.), *Nonextensive Entropy—Interdisciplinary Applications*, SFI Working paper 2003-12-068, Santa Fe Institute: Oxford University Press, 2003, 387–98.

Godwin D, Barry RL, Marois R. Breakdown of the Brain's functional network modularity with awareness. *PNAS* 2016;**112**: 3799–804.

Grau C, Escera C, Yago E, *et al*. Mismatch negativity and auditory sensory memory evaluation: a new faster paradigm. *NeuroReport* 1998;**9**:2451–6.

Grunwald P, Vitanyi P. Shannon Information and Kolmogorov Complexity. arXiv:cs/0410002, 2004.

He BJ. Scale-free brain activity: past, present, and future. *Trends Cogn Sci*, 2014;**18**:480–87.

He S, Wang H, Jiang ZH, (2014). Identifying user behavior on twitter based on multi- scale entropy. In: *2014 IEEE Int. Conf. on Security, Pattern Analysis, and Cybernetics (SPAC)*.

Hofstadter DR. *I Am a Strange Loop*. New York: Basic Books, 2007.

Hohwy J. *The Predictive Mind*. Oxford: Oxford University Press, 2013.

Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Netw* 1991;**4**:251–7.

Hudetz AG, Liu X, Pillay S, *et al*. Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neurosci Lett* 2016;**628**:132–5.

Hutter M. Algorithmic information theory. Scholarpedia, 2007;**2**:2519, doi: 10.4249/scholarpedia.2519.

Kanoh S, Futami R, Hoshimiya N. Sequential grouping of tone sequence as reflected by the mismatch negativity. *Biol Cybern* 2004;**91**:388–95.

Kaspar F, Schuster HG. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys Rev A* 1987;**36**:842–8.

Kayama Y. Complex networks derived from cellular automata. arXiv:1009.4509, 2010.

Koch C, Massimini M, Boly M, *et al*. Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci* 2016;**17**: 207–321.

Koch C, Tononi G. Can machines be conscious? *Spectrum* 2008;**45**: 54–59.

Kovács I, Papathomas T, Yang M, *et al*. When the brain changes its mind: interocular grouping during binocular rivalry. *PNAS* 1996;**93**:15508–11.

Lagercrantz H, Changeux J-P. Basic consciousness of the newborn. *Semin Perinatol* 2010;**34**:201–6.

Laureys S. The neural correlate of (un)awareness: lessons from the vegetative state. *Trends Cogn Sci* 2005;**9**:556–9.

Lempel A, Ziv J. On the complexity of finite sequences. *IEEE Trans Inf Theory* 1976;**22**:75–81.

Li M, Vitanyi P. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer, 2008.

Li W, Nordahl MG. Transient behavior of cellular automaton rule 110. Technical Report SFI Working paper: 1992-03-016, Santa Fe Institute, 1992.

Lloyd S. The computational capacity of the universe. *Phys Rev Lett* 2002;**88**:1–4.

Mainzer K, Chua L. *The Universe as Automaton: From Simplicity and Symmetry to Complexity*. Berlin Heidelberg: Springer, 2012.

Manor B, Costa MD, Hu K, *et al*. Physiological complexity and system adaptability: evidence from postural control dynamics of older adults. *J Appl Physiol* 2010;**109**:1786–91.

Massimini M, Ferrarelli F, Huber R, *et al*. Breakdown of cortical effective connectivity during sleep. *Science* 2005;**309**: 2228–32.

Mhaskar H, Liao Q, Poggio T, (2016). Learning functions: when is deep better than shallow. Technical Report CBMM Memo No. 045, CBBM.

Mnih V, Kavukcuoglu K, Silver D, *et al*. Human-level control through deep reinforcement learning. *Nature* 2015;**518**:529–33.

Lambert MV, Sierra M, M. P, David A. The spectrum of organic depersonalization. A review plus four new cases. *J Neuropsychiatry Clin Neurosci* 2002;**14**:141–54.

Näätänen R, Gaillard AW, Mäntysalo S, *et al*. Early selective-attention on evoked potential reinterpreted. *Acta Psychol Amst* 1978;**42**:313–29.

Ninagawa S. Computational universality and 1/*f* noise in elementary cellular automata. In: Ninagawa S (ed), *22nd International Conference on Noise and Fluctuations (ICNF)*, number 10.1109/ICNF.2013.6578934 in International Conference on Noise and Fluctuations (ICNF), 2013.

Nishimoto S, Vu AT, Naselaris T, *et al*. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr Biol* 2011;**21**:1641–6.

Pusiol G, Esteva A, Hall SS, *et al*. Vision-based classification of developmental disorders using eye-movements. In: *MICCAI2016*, 2016.

Quiroga Q, Reddy L, Kreiman G, *et al*. Invariant visual representation by single neurons in the human brain. *Nature* 2005;**435**: 1102–7.

Ravasz E, Barabasi A-L. Hierarchical organization in complex networks. *Phys Rev E* 2003;**67**:1–7.

Ray C, Ruffini G, Marco-Pallarés J, *et al*. Complex networks in brain electrical activity. *Eur Phys Lett* 2007;**79**:38004.

Reggia JA. The rise of machine consciousness: studying consciousness with computational models. *Neural Netw* 2013;**44**: 112–31.

Ruffini G. Information, complexity, brains and reality ("Kolmogorov Manifesto"). http://arxiv.org/pdf/0704.1147v1, 2007.

Ruffini G. Reality as simplicity. arXiv: https://arxiv.org/abs/0903.1193, 2009.

Ruffini G. Models, networks and algorithmic complexity. *Starlab Technical Note*—arXiv:1612.05627, TN00339 (doi:10.13140/RG.2.2.19510.50249), 2016.

Schartner M, Seth A, Noirhomme Q, *et al*. Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. *PLoS One* 2015;**10**:1–21.

Schartner MM, Carhart-Harris RL, Barrett AB, *et al*. Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Sci Rep* 2017;**7**:46421.

Seager W, Allen-Hermanson S. *Panpsychism*, The Stanford Encyclopedia of Philosophy. Stanford University: Metaphysics Research Lab, 2015.

Seth AK. Interoceptive inference, emotion, and the embodied self. *Trends Cogn Sci* 2013;**17**:565–73.

Seth AK. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn Neurosci* 2014;**5**:97–118.

Seth AK. *The Real Problem*. Aeon. Co., 2016. https://aeon.co/essays/the-hard-problem-of-consciousness-is-a-distraction-from-the-real-one.

Shic F, Chawarska K, Bradshaw J, *et al*. Autism, eye-tracking, entropy. In: *7th IEEE International Conference on Development and Learning, 2008, ICDL 2008*. Monterey, CA, USA, DOI: 10.1109/DEVLRN.2008.4640808.

Siegelmann HT, Sontag E. On the computational power of neural nets. *J Comput Syst Sci* 1995;**50**:132–50.

Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos Trans R Soc Lond B Biol Sci* 2009;**364**:3549–57.

Soler-Toscano F, Zenil H, Delahaye J-P, *et al*. Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLoS One* 2014;**9**

Stanley GB, Li FF, Dan Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J Neurosci* 1999;**19**:8036–42.

Stiefel KM, Merrifield A, Holcombe AO. The Claustrum's proposed role in consciousness is supported by the effect and target localization of *Salvia divinorum*. *Front Integr Neurosci* 2014;**8**: 20.

Taylor P, Hobbs JN, Burroni J, *et al*. The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Sci Rep* 2015;**5**:1–18.

Tononi G, Boly M, Massimini M, *et al*. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.

Tononi G, Koch C. The neural correlates of concsciousness—an update. *Ann N Y Acad Sci* 2008;**1124**:239–61.

Turing AM. On computable numbers, with an application to the Entscheidungsproblem. *Proc Lond Math Soc* 1936;**2**:230–65.

Uhrig L, Dehaene S, Jarraya B. A hierarchy of responses to auditory regularities in the macaque brain. *J Neurosci* 2014;**34**: 1127–32.

Van Gulick R. Consciousness. In Zalta EN (ed.), *The Stanford Encyclopedia of Philosophy*, Stanford University: Metaphysics Research Lab, 2016.

West GB. The origin of universal scaling laws in biology. *Phys A* 1999;**263**:104–13.

Wolpert D. The physical limits of inference. *Phys D* 2008;**237**: 1257–81.

Yu K, Blake R. Do recognizable figures enjoy an advantage in binocular rivalry? *J Exp Psychol Hum Percept Perform* 1992;**18**: 1158–73.

Zenil H, Kiani NA, Tegnér J. (2015a) Methods of information theory and algorithmic complexity for network biology. *Seminars in Cell & Developmental Biology*, 2016;**51**:32–43.

Zenil H, Soler-Toscano F, Dingle K, *et al*. Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks. *Physica A: Statistical Mechanics and its Applications*, 2014;**404**: 341–58.

Ziv J, Lempel A. Compression of individual sequences by variable rate coding. *IEEE Trans Inf Theory* 1978;**IT-24**:530–6.