An algorithmic information theory of consciousness (annotated)

Supplementary Material — Neuroscience of Consciousness, 2017, 3(1): nix019

Giulio Ruffini

Starlab Barcelona, Avda. Tibidabo 47bis, 08035 Barcelona, Spain and Neuroelectrics Corporation, 210 Broadway, Cambridge, MA 02139, USA, giulio.ruffini@starlab-int.com

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).

Keywords: 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

1. 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 model their input/output fluxes of information (I/Os) with simplicity as a fundamental driving principle. (Ruffini, 2007, 2009). Furthermore, we argue, 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[1].
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, compressors of information for survival. Cognition and phenomenal consciousness arise from modeling, compression and data tracking using models. At this stage, from

[1] Consider the following *gedanken (G1)* (thought experiment). A subject is connected to a very advanced, fully immersive VR system capable of controlling all the subject's I/Os. Here the experience may be mediated using peripheral means, or perhaps the I/Os are patched directly to the subject's CNS using an invasive technology. The I/Os are relayed through a high speed optical fiber to a computer that interacts (interprets outputs and sends back inputs to the subject) so that she feels to be laying in a beach in Hawai'i—with, let us assume, the experience feeling completely real. In this hypothetical scenario, which while not possible today certainly seems technologically feasible in the future, the universe experienced by this subject is fully described by the information carried by the optical fiber—the actual sequence of bits. It is in this sense that we may say that as far as our subject is concerned, "there is information", and the rest is inference.

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)[2]. 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, avalanches, etc.).
5) We return to the subjective and hypothesize that structured, graded, 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

[2]Here we aim to provide a concrete mathematical framework to study AIT. However, we note that CA-based physical theories are also being studied ('t Hooft, 2014).

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 closing section we discuss some limitations and open questions.

2. Computation, compression and cognition

The definition of a Universal Turing Machine (Turing (1936)) provides the mathematical foundation for computation and algorithmic information theory, and hence plays a key role in KT. Although our starting point is mathematical[3], it is readily linked to physics[4]. In practice, all formulations of fundamental physics theories can be

[3] In another *gedanken (G2)*, suppose that someday we develop algorithms for artificially intelligent systems that pass the Turing test—that we all come to agree are conscious. If so, we would be implicitly asserting that consciousness is truly a mathematical phenomenon that, through computation, some algorithms possess. We could indeed launch the same algorithms in different types of computers with identically results, for example. In this way, we would abstract away the potential for conscious experience from its instantiation in matter, just as with computation.

[4] In KT, physical objects, including physical computers, as well as physical theories, are emerging phenomena: models created by cognitive systems to compress data. Mathematics, on the other hand, although it is also constructed by cognitive systems, could be argued to exist in a Platonic sense. We may summarize our reasoning as follows: 1) *Computation is mathematical:* Computation is a mathematical concept (Turing defined it as a mathematical structure); 2) *All physical things are mathematical:* all I have access to is information, and through mathematics (and computation) I can create models to simplify this information, and physical concepts (matter, energy, chairs) emerge in that way, as models; 3) *The universe is, at root, mathematics:* from the above, we come to the more fundamental hypothesis that all the data we get originates from something we can call the universe: a mathematical structure, a grand model (which rides on the concept of computation). Thus, as far as *I* can tell, the universe is a mathematical structure. It may be that the data I receive is actually generated by something else (physical?), but I don't have access to that and never will. While this may sound a bit extreme, it helps to imagine it at first as being reasoned by a freshly created AI system rather than a human brain, who gets all its I/Os from an optical link—as in Jaynes (2003)'s robot.

The computation-physics link is based on the Church-Turing conjecture. It loosely states that any physically-realizable "effective procedure" (a calculation stemming from the dynamics of some physical system—including those carried out by living beings, which are seen to implement mechanistic procedures), can be translated into an equivalent computation involving a TM (Copeland, 2015), or, equivalently, using the lambda calculus, general recursive functions, cyclic tag systems or recurrent neural networks (RNNs). The Church-Turing thesis is complemented by the principle of computational equivalence: *"almost all processes that are not obviously simple can be viewed as computations of equivalent sophistication"* ((Wolfram, 2002), pp. 5 and 716-717). More specifically, the principle states that systems found in the natural world can perform computations up to a maximal ("universal") level of computational power, and that most systems do in fact attain this maximal level of computational power. Consequently, *"most systems are computationally equivalent. For example, the workings of the human brain or the evolution of weather systems can, in principle, compute the same things as a computer. Computation is therefore simply a question of translating inputs and outputs from one system to another."* Of course, such capability is not equivalent to realization. The programs that run

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 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 then the conjecture (called "digital physics") that the universe is discrete and isomorphic to a TM[5]. 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 are examples of TMs which may be appropriate at different levels of description[6].

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: a) they evolve according to local homogenous rules and b) 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[7]. 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

---

on such systems are crucial.

There is no proof for these principles linking the physical world and mathematics, but every realistic model of computation discovered so far has been shown to be equivalent and we seem to be able to model the world using computation.

For completeness, we mention a new approach to the study of computing and complexity. An absolute measure of complexity using the concept of (strong) *inference machines* has recently been proposed (Wolpert, 2008). Inference machines theory (IMT) is a mathematical formalization of physical devices performing observation, prediction or recollection tasks—what we call here cognitive systems. We could use this formalism as a starting point equally well.

[5]This idea builds from early work by Jaynes (Jaynes, 1957), John Wheeler ("it from bit" (Wheeler, 1990)), Zuse (Zuse, 1967), Fredkin (Fredkin, 2003, 2004), Wolfram (Wolfram, 2002), Tegmark (Tegmark, 2014b) and others such as Gerard 't Hooft ('t Hooft, 2014).

[6]E.g., a brain can be seen as a RNN embedded in a CA universe (small scale description vs. coarse grained one). Both of them are TMs. Mathematically, the entire discussion could take place talking only about abstract TMs, but this description makes the connections with physics and neuroscience clearer.

[7]Interestingly, research within other information-theoretic frameworks for consciousness (IIT) has used CAs to model the emergence of adaptive systems (Albantakis and Tononi, 2015).

(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; Ninagawa, 2013; Mainzer and Chua, 2012). 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[8].

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,

[8]See for example Rendell (2014). The physics—or rather, mathematics—we describe are embedded in spacetime. We can imagine for simplicity a 2D board (CA) representing the Game of Life—the first implementation of a Turing Machine in a CA, Conway, (Gardner, 1970)—as it evolves in time. The rules of this game (laws of physics) describe how a 2D spatial slice is to be computed from the prior one. Note that the very concept of computers requires, in a sense, spacetime: space is what contains the tape/program and state machine, time describes the evolution of computation. We can thus picture programs as evolving in time as they are executed. On the other hand, in principle, programs (models) live on spatial slices (e.g., the tape on a TM, or the state of a 2D CA at any given time). Models are extended structures in CA space (such as a *Blinker* or *Pulsar* in the Game of Life). We can compare such spatially extended structures to other simpler ones that transport information across the CA (information carriers such as *gliders*) and require a spacetime description.

Starting from such simple CA models (yet ones capable of universal computation), we can illustrate the main KT ideas in a somewhat extreme Platonic way. Consider a toy model in which the universe is described by the rules of cellular automaton such as Conway's The Game of Life or Wolfram's elementary Cellular Automata Rule 110 (Wolfram, 2002; Cook, 2004), both of which are known to implement universal computation (see (Berto and Tagliabue, 2012) for a lucid discussion of CAs). A brain in this context corresponds to a specific configuration in which a subset of CA cells perform computation and prediction, and, for example, adapt (via replication and natural selection) to dodge harmful patterns (bullets) in the CA. Thus, seen from the "outside" the brain is just a pattern of changing states in the CA, but quite a special one (in a way that can be mathematically characterized). We would call this CA a brain or agent, and its goal would be to survive as many CA cycles as possible. What is even more interesting, the existence of such a pattern is intrinsic to the CA rules and initial conditions. The picture just described is succinctly summarized by the CA equations plus the initial conditions used to run it. In a sense, the CA need not be even run: once defined, it implicitly specifies a spacetime of states, a history which is in a real sense static (time is a model (Barbour, 1999)). In a CA universe model, information could be transported by some emergent CA particle-like structures that interact (are entangled with) with others in the CA world. Information transmission here just means that the state of a system can affect the state of another by the CA dynamics in a specific way. In this view, in this emergent definition of (meta) information, we are aligned with (Rovelli, 2015), which postulates that information is a meaningful *emerging concept* as the mutual information between two interacting systems, or with (Deutsch and Marletto, 2014), where information is seen to be a derived concept from computation. Here we further emphasize the role of the mutual *algorithmic* information between systems. Just as in Conway's Game of Life, CA structures/patterns such as CA-bullets, CA-brains or CA-photons are all meta-structures produced by the CA. It is in this way that we define the concept of algorithmic mutual information between structures. For example, the emission and then absorption of a CA-photon in two CA-structures will increase their mutual information. We will return to CAs below in an effort to link compressive agents with measurable metrics.

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)[9]. 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).

2.1. 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" (Figure 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—neural networks 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. E.g., 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.,

[9] A hierarchical network can achieve the same as a shallow one, but, in a simpler way, using less nodes. By a *compositional function* we mean, e.g.,

$$f(x_1, ..., x_d) = h_1(h_2(...(h_j(h_{k1}(x_1, x_2), h_{k2}(x_3, x_4)), ...)),$$

which is best visualized as a compositional tree. A neural network can be seen as implementing a program to compute a function in a sequence of steps in a programming language where each layer represents a set of parallel computational steps in a sequence. If the function to be computed is simple (allows for a short description), it is fairly intuitive that deep networks can provide more succinct, and hence easier to train network structures than shallow ones, which are limited to single step computations (with a smaller programming language repertoire).

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 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[10]. 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.

Based on 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 outputs streams, which also feedback to the model[11]. A classical thermostat or a machine learning classifier is not an agent by this definition, but new artificial intelligence systems being developed are. As an

[10]This is especially true if the agent's actions have a great impact on its immediate environment. That is, self-modeling is important in proportion to the impact the agent has in its input stream.

[11] According to KT, the key elements in Figure 1 must exist in some form in a cognitive system such as the brain. Note that this must be seen as a rough initial approximation to the problem. E.g., we did not explain how the model has been discovered, only that one is available, perhaps assembled from available ones by integrating and/or tweaking them, and that it is malleable (learning can take place). In fact, we can allow for more than one model being available in a probabilistic sense: there may be several that explain the data with similar performance (i.e., posteriori probability, taking into account both data matching and length of description). This leads naturally to the idea of probabilistic inference, which in the end is itself a model. Recall that compression is a tool for model discovery taking into account model match with data, since the compressed length is a function of both the kernel of regularities (the sufficient statistic of Kolmogorov) and the uncompressible part, which we can think of as error. The Solomonoff prior provides a probability associated with each model. The action module

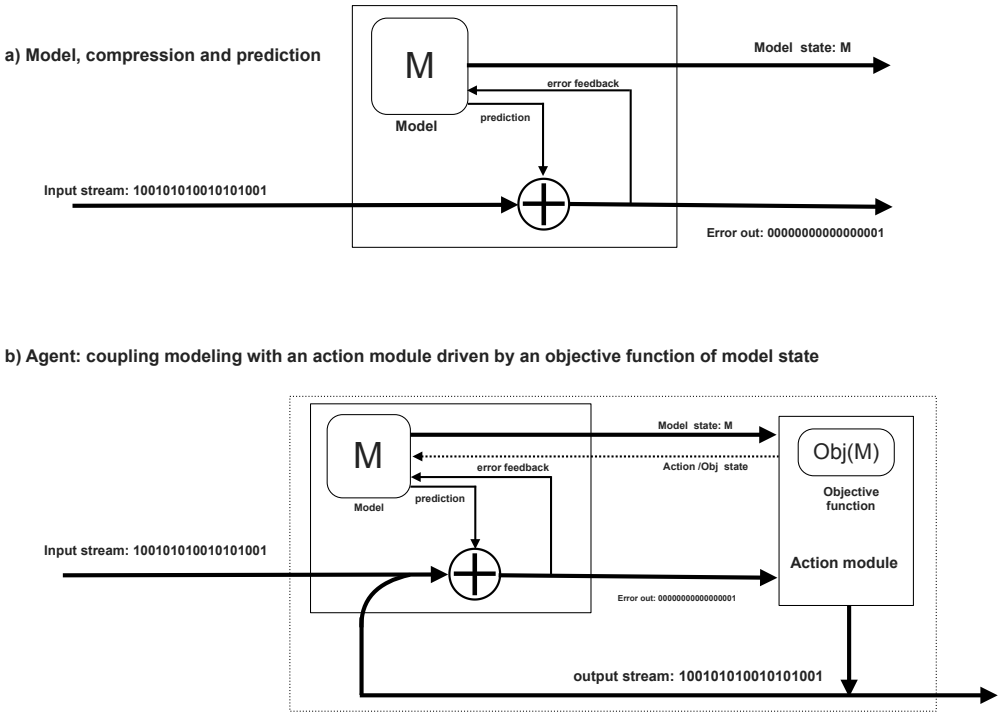

Figure 1: Top (a): Modeling for predictive compression. The arrows indicate information flow, and the circled plus sign comparison/difference (XOR gate). Bottom (b): An agent with coupled modeling and an action modules. The action module contains an optimization objective function (e.g., homeostasis), 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.

must make a choice based on this information, through the objective function. Each model choice will translate into a future action and a consequent future forecast of objective function change. The task of the action module is to choose the outputs ("action") that in probability, will lead to the greater positive change in the objective function, very much as a policy needs to be solved for in reinforcement learning (Mnih et al., 2015). The Action module will rely on the Model to optimize its policies.

The objective function can be seen as assigning valence, positive, negative or null (one dimensional scale), driving the agent to homeostasis. Evolution and natural selection will strongly shape such optimization functions.

We also make note here of the *boundary problem* of consciousness, and here of agents (Fekete et al., 2016): how to mathematically define the boundary, the "information membrane" of an agent. The concept of mutual algorithmic information of agent and the world may be a useful starting point. In order to attempt to define a boundary of an agent, we can refer to the CA picture of an agent in the universe (the world). Formally, the agent is a subset $\mathcal{A}$ of the CA with a very high mutual algorithmic information (MAI) with the world. We may like to think of $\mathcal{A}$ as a connected set in the CA, but this is

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)[12] .

2.2. Simplicity and Kolmogorov Complexity ($\mathcal{K}$)

Compression (and therefore simplicity) were formalized by the mathematical concept of algorithmic complexity or *Kolmogorov complexity* ($\mathcal{K}$), co-discovered during the second half of the 20th century by Solomonoff, Kolmogorov and Chaitin[13]. 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[14]. An associated useful notion is the *mutual algorithmic information (MAI)* between two strings (Grunwald and Vitanyi, 2004), the algorithmic analog of Shannon mutual information[15].

---

not actually necessary. A potential definition of a minimal agent may make use of a quantity such as the mutual algorithmic information per cell of the proposed agent, as follows: *Def:* A *minimal agent* is a minimal subset $\mathcal{A}$ of the CA with high effective MAI per cell with the CA: removing one or more cells from $\mathcal{A}$ lowers its MAI/cell with the CA, and adding cells does not increase it MAI/cell significantly. (Chaitin, 1991; Ruffini, 2007, 2009). Using such a definition we may attempt to identify local maxima of MAI that may be associated with structured experience in agents. However, such boundaries will probably be fuzzy. This is not a problem in KT, since it is an inherently panpsychist theory, where consciousness is not a binary quantity restricted to special arrangements of matter. "Panpsychists see themselves as minds in a world of mind." (Wikipedia, but see also Seager and Allen-Hermanson (2015)).

[12] In fact, the model used in Deep Reinforcement Learning is essentially equivalent to the one described here for an Agent. Our Action module is the one implementing a policy, which needs to be optimized using the Model. Having access to a succinct model simplifies drastically the search for an optimal policy. An interesting point is that the Model can only be as rich as the universe it is trying to model. The analogy is that human conscious level has an upper bound in the algorithmic complexity of the universe.

[13] Kolmogorov complexity is also known as 'algorithmic information', 'algorithmic entropy', 'Kolmogorov -Chaitin complexity', 'descriptional complexity', 'shortest program length' and 'algorithmic randomness'.

[14] That is, if $\mathcal{U}$ is a universal computer, then for any computer $\mathcal{A}$ and all strings x we can easily show that $K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$, where the constant $c_{\mathcal{A}}$ does not depend on the string x (Cover and Thomas, 2006).

[15] Let $E_X[Q(X)]$ denote the expectation value of some random variable $Q(X)$ under a probability distribution $P(X)$. We recall first that the mutual (Shannon) information (Cover and Thomas, 2006) between random variables X and Y is

$$I(X;Y) = H(Y) - H(Y|X),$$

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 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 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[16]. 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

where $H(X)$ is the entropy of X, $H(X) = -E_X[\log(P(X)]$, and $H(Y|X)$ is the entropy of Y conditioned on X,

$$H(Y|X) = -E_{X,Y}[\log P(Y|X)] = H(Y,X) - H(X).$$

Analogously, the mutual algorithmic information (MAI) between two strings x and y is

$$I_{\mathcal{K}}(x:y) = \mathcal{K}(y) - \mathcal{K}(y|x),$$

where $\mathcal{K}(y|x)$ is the complexity of the string y if the computer has access to x (Li and Vitanyi, 1997; Grunwald and Vitanyi, 2004). In practice, we can use the relation $\mathcal{K}(x,y) \approx \mathcal{K}(x) + \mathcal{K}(y|x)$ so that $K(y|x) \approx K(x,y) - K(x)$ and

$$I_{\mathcal{K}}(x:y) \approx \mathcal{K}(y) + \mathcal{K}(x) - \mathcal{K}(x,y).$$

As a proxy for algorithmic complexity, LZW can be used to estimate MAI using this equation.

[16] See for example the concept of *Universal (Levin) Search*, which resolves the problem of computability of $\mathcal{K}$ and adds an element of practical relevance by penalizing slow programs (Hutter, 2007), or *Logical Depth* (Bennet, 1988; Zenil et al., 2015b), where the complexity of a string is defined by the time that a computation process takes to reproduce the string from its shortest description. However, note that these definitions build on Kolmogorov complexity rather than supersede it in a fundamental way. Another approach is to limit the programming language repertoire. E.g., limiting the search to primitive recursive functions, a subset of recursive functions, avoids the halting problem in algorithmic complexity at the expense of giving up universal computation (Ruffini, 2016). LZW is a prime example of this.

$\mathcal{K}$[17].

The connection between simplicity, statistics and prediction was developed by Solomonoff through the definition of the *algorithmic or universal probability* $P_{\mathcal{U}}(x)$ *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 $P_{\mathcal{U}}(x) \approx 2^{-\mathcal{K}_{\mathcal{U}}(x)}$. 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 is generated by ferrets typing programs or by other random program generating mechanisms[18]. 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:

[17] The main idea in LZW is to look for repeating patterns in the data, and instead of rewriting repeating sequences, refer to the last one seen Lempel and Ziv (1976). As Kaspar and Schuster (1987) clearly state, LZW is the Kolmogorov Complexity computed with a limited set of programs that only allow copy and insertion in strings.

> "We do not profess to offer a new absolute measure for complexity which, as mentioned already, we believe to be nonexistent. Rather, we propose to evaluate the complexity of a finite sequence from the point of view of a simple self- delimiting learning machine which, as it scans a given n- digit sequence $S = s_1 \cdot s_1 \cdot ...s_n$, from left to right, adds a new word to its memory every time it discovers a substring of consecutive digits not previously encountered. The size of the vocabulary, and the rate at which new words are encountered along S, serve as the basic ingredients in the proposed evaluation of the complexity of S."

Entropy rate, given by $H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, ..., X_n) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}..., X_1)$ and LZW are essentially equivalent for stationary stochastic processes (Cover and Thomas, 2006).
LZW can be used to obtain an upper bound on $\mathcal{K}$, but, given its reduced programming repertoire, it will fail to compress random-looking data generated by simple, but highly recursive programs, e.g., the sequence binary digits of π, with more sophisticated regularities than sequence repetition (deep programs (Ruffini, 2016)). As an example of how LZW or entropy rate are limited tools for compression, and therefore poor cousins of algorithmic complexity, consider a string and its "time-reversed" version in a file with the ordering of symbols inverted, or a string and "time-dilated" string where each symbol is repeated, say 2 times. Such simple algorithmic manipulations will not be detected and exploited by LZW. It is interesting to note that the expectation value of algorithmic complexity of a *stochastic* process is proportional to its entropy rate as well (Cover and Thomas, 2006). However, all processes in nature are probably deterministic, not stochastic.

[18] The precept that short explanations are in this sense more likely is the essence of the Minimum Description Length and the Minimum Message Length approaches to statistical inference (Li and Vitanyi, 2008; Wallace and Dowe, 1999; Vitányi and Li, 2000). Among possible explanations of an observed data string, those which are shortest (program plus error) are more likely. Similarly, in machine learning one speaks of regularization, sparsity or other approaches for reducing the learning problem to simpler (smaller) spaces of solutions.

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.

3. 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[19].

3.1. 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* [20]. It is ultimately limited by the algorithmic complexity of the universe the agent is in and the resources it has access to.

Returning to Figure 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

[19] It seems reasonable to assume that being conscious of something will be shaped by what that something actually is, which we equate to a model here.

[20] This is a different meaning than, e.g., the one used in neurology:*level of consciousness is a measurement of a person's arousability and responsiveness to stimuli from the environment.*

for[21], 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)[22].

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, for example, represents such a system[23]. 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[24]. 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.

3.2. 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 algo-

[21] Models are thus "eternal", invariants. Of course, as invariant rules they may specify rules for dynamic changes in parameters or sub-rules themselves.

[22] To clarify this point, consider another *gedanken (G3)*. If we place somebody in a dark room, the subject will receive a simple stream of data, consisting of, say, mostly 0s (it's dark and silent). Now, what model is this person to make of the data stream? We need to keep in mind that as a modeling tool the brain has access to this data as well as all other past experience. What model will it come up with? Let's say that it is "I have been placed in a dark room". But is this a simple model? If the recent experience is compatible with this (e.g., the person has been led by a friend with eyes closed and it is our subject's birthday) the model will actually be "I have been placed in a dark room by my friend, because I am about to experience a birthday party!". This model fits well with all the available data, not just the string of recent 0s. It is a good model, and the subject will feel the experience as being very real. If however, our subject were driving on the highway and suddenly a demigod from the Enterprise teleports her to a dark room, her "reality meter" will be lower according to KT. The event will feel unreal, because no simple model will be readily available to match the data.

[23] In fact, the distinction between life, cognition and consciousness is as a matter of degree in KT.

[24] As an applied research field, Presence studies how to produce real-feeling experiences in mediated interaction—the phenomenon of behaving and feeling as if a mediated experience were real (Sanchez-Vives and Slater, 2005).

rithmic 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[25].

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[26] (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[27], and we may call them *rare sequences*. Healthy, conscious brains should produce such data. And although providing improved bounds on algorithmic complexity remains a challenge,

[25]The reason for this is that although the models for the world are simple they generate data with high entropy rates. Wolfram points out that such functions/models are rather common (Wolfram, 2002). A brain tracking such data streams will generate complexity itself, mirroring what happens in the world.

[26]Certainly, further research is needed to establish connections between algorithmic and other flavors of complexity such as multiscale entropy (MSE) (Costa et al., 2002), entropy rate or power laws.

[27]The probability that a random sequence can be compressed by more than k bits is no greater than 2^{-k} Cover and Thomas (2006)

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 is 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.

3.3. Mutual algorithmic information 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[28].

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 mutual algorithmic information $I_{\mathcal{K}}(x : y)$ 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

[28] E.g., consider an experiment in which an image of a face or a scene is presented to a subject. Suppose we find a neuron that fires exclusively when seeing a particular face regardless of orientation or lighting, and we show the subject a sequence of faces and other object images. Now, the mutual algorithmic information between neuronal activity and images is high (we can use knowledge of the latter to compress the former, or vice-versa). However, what is notable is that neuronal activity contains this information in compressed form (not as a sequence of raw bits encoding pixels). Let us imagine now that this is also an fMRI experiment, and that we can, with this data alone create a machine learning classifier that outputs the class of the image that has been presented from fMRI voxel activations. Can we conclude from this that the subject is conscious of the image? KT would say that as long as the representation is compressed, structured experience will occur. This will happen, for example, in the "grandmother" cell example, but also if there is a specific pattern of activation associated to the event that responds in some invariant way.

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[29].

3.4. Relation to Integration Information (IIT), Global Workspace (GWT) and Predictive Processing (PP) 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

[29] Consider the question whether there is useful, compressed information about world data x in a brain or agent with data y. E.g., x is a sequence of images (pixels) of a face or a cat, one billion of them, and y is the data stream from a face cell that fires when it sees a face. Or the negativity signal in a MMN paradigm with a billion trials. The following two conditions need to be met to assert that there is useful, compressed information about the world in a brain or agent:

1- Knowing brain data x_b means we know a lot (let's say all) about some given world data x:

$$\mathcal{K}(x|x_b) = \mathcal{K}(x) - I_{\mathcal{K}}(x : x_b) \approx 0$$

So

$$\mathcal{K}(x) \approx I_{\mathcal{K}}(x : x_b)$$

or

$$I_{\mathcal{K}}(x : x_b)/\mathcal{K}(x) \approx 1$$

That is, the brain knows about the world.

2- Moreover, x_b is a good compressed version of world data x:

$$l(x_b) << l(x), \quad l(x)/l(x_b) >> 1$$

Putting it all together, a necessary condition is that

$$\frac{I_{\mathcal{K}}(x : x_b)}{\mathcal{K}(x)} \frac{l(x)}{l(x_b)} > 1$$

which we may rewrite as

$$\frac{I_{\mathcal{K}}(x : x_b)}{l(x_b)} \frac{l(x)}{\mathcal{K}(x)} > 1.$$

The first factor requires that there is high MAI and it is in compressed form in the agent (to be ≈1), the second that world data be non-obvious (not yet compressed). Note, however, that sensory streams are disconnected from the agent, 2 follows if we assume a reasonably small memory capacity in the agent—information then must be compressed.

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[30]. If structured experience is shaped by models, our belief in a particular model (as driven by the input/output 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[31]. 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[32]. We note that simple programs (in the limit of Kol-

[30] On the surface, the theories use different mathematical frameworks. As it has happened in other occasions, it may be that they are actually equivalent. Regarding phenomenal structure, KT relates to the IIT axioms in the following ways:

- Experience exists (intrinsically, independent of external observers). In KT we start from the same assumption.
- Composition. Experience is structured (it has many aspects): so is a model that integrates multiple sources of data to provide a unified, organized experience.
- Information. Experience is differentiated (one out of many): it is what it is by differing in its particular way from many others: so is a model that has been selected against other in terms of its merits as the simplest and best data fitting (i.e., compressive).
- Integration. Experience is unified (it is "one"): it cannot be reduced to non-interdependent components: a model does this automatically, as a simple program that cannot be broken into parts.
- Exclusion. Experience is unique (it is only one), in content and spatiotemporal grain: it is not a superposition of multiple experiences, with less or more content, flowing at faster or slower speed at once. Selection of the best model for available data eliminates ambiguities.

[31] As an example of IIT's *information*, *integration* and *differentiation* in KT, let us return to CA Rule110 with some arbitrary initial conditions. Consider the "time series" associated to a couple of columns in the resulting spacetime diagram. Each series will be highly entropic (highly informative), as well as differentiated (the two series will be uncorrelated). Yet, the entire picture is highly integrated, bound by the underlying simple model plus initial conditions.

[32] KT suggests as well that a possible future version of IIT may consider algorithmic information as opposed to Shannon information derived concepts (as, e.g., used in the definition of C_n or φ, where the core elements are probability distributions and their distance). An interesting mathematical line of research to compare the theories is to study systems capable of universal computation running simple models in terms of IIT's "neural complexity" C_n or its generalization φ, which describe information integration across variables or nodes. For example, a cellular automaton running a simple model will presumably display strong integration as computed using φ or C_n, as recently discussed in Albantakis and Tononi (2015). Neural networks offer another interesting scenario for such studies.

mogorov) 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 neural networks 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)[33].

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 which integrates information from auditory, visual and somatosensory information, as well as from the thalamus and limbic system[34] (Koch et al., 2016).

[33]The core idea is that an efficient, simple network activated by an input which it recognizes will be more sensitive to perturbations of its nodes. In a simple network running a deep function (recursive, Kolmogorov simple), all nodes play a crucial role in a pattern tracking task.

[34]The TPJ has been identified as a candidate for the generation of Presence qualia and proposed as a full neural correlate of consciousness. Temporo-parietal areas are also associated with the generation of out-of-body experiences through electrical (Blanke et al., 2002) or TMS (Blanke et al., 2005) stimulation of the TPJ, or in patients with lesions in that area.
A condition of interest in this respect is called "derealization" (an alteration in the perception or experience of the external world so that it seems strange or unreal) and "depersonalization" (alteration in the perception or experience of the self so that one feels detached from, and as if one is an outside observer of, one's mental processes or body), can appear following temporal lobe epilepsy, migraine or head injury (Sierra et al., 2002; MV Lambert and David, 2002). Left-sided temporal lobe dysfunction and anxiety are suggested as factors in the development of depersonalization. Hemineglect is another

Predictive processing theory (PP) (Friston, 2009; Hohwy, 2013; Seth, 2013; Clark, 2013; Seth, 2014) is also closely related to KT, with a focus on the predictive efficiency afforded by simple models of I/Os. It maintains that in order 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 which 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.

4. Experimental methods in KT

4.1. 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 Slater (2009); Ruffini (2009)). As we progress higher in the modeling hierarchy, Bayesian prior expectations[35] 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

example (Husain, 2008) of relevance, although the view that damage to parietal areas, including the TPJ, is responsible has been challenged.

Another candidate location for integration is the *claustrum* (Crick and Koch, 2005; van den Heuvel and Sporns, 2013; Koubeissi and Bartolomei, 2014; Stiefel et al., 2014; Fischer et al., 2016), a thin sheet of grey matter below the insular cortex with bidirectional connections to a number of cortices, including the primary motor, premotor, prefrontal, auditory, and visual, among others, combining multiple information modalities. Finally, we mention the insular cortex as well, which is involved in multimodal synchrony checking (Bushara et al., 2001) and also pain and emotion.

[35] Bayesian modeling emerges here as part of the modeling, compressing process.

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. E.g., 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, for example, 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 and 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[36] (Bekinschtein and 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 (for example, Benidixen and Schröger (2008) discuss how rapidly simple but abstract rules are learned)[37]. This is also addressed in Bekinschtein and et al. (2009)

[36] As discussed in (Bekinschtein and et al., 2009), unlike the P300b, neither MMN nor P300a have been associated empirically with conscious access.

[37] The oddball paradigm could be expanded in a multi-modal manner (visual, somatosensory, auditory, etc.) using immersive technologies (VR) to fully manipulate interaction (I/Os). If an agent is capable of constructing a simpler model for data it will allow the agent to store it better (less space needed) and also to predict the future more precisely. The experiments would involve 1) creating simple world models that we can manipulate, and generate data streams from them, 2) have subjects

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[38]. 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.

4.2. 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 neural networks, 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[39]. In addition, various non-invasive stimulation

in different states of consciousness try to track these I/Os in a task in which they have to remember or predict future events (to measure if they actually identified a good model for the I/Os), 3) use subjective reports of presence to assess experience and also which models they were actually using for the experience. Availability and access to simple models should map into more real feeling VR experiences—as we argued happens in the rubber arm experiment (Ruffini, 2009).

[38] The rules are simple and we can actually compute their algorithmic complexity. The local rule is simply coded as "repeat n-times the tone X" (as in $R_n(X)$), while the global rule requires the specification of a sequence of several tones (e.g., $R_n(XXXX)$ or $R_n(XXXY)$). More complex rules could be explored, of course.

[39] As already mentioned, despite being a good starting point for the study of $\mathcal{K}$, LZW is limited. Machine learning may become a key enabling technology to provide tighter upper bounds for $\mathcal{K}$.

methods, such as transcranial current stimulation (tCS) can be used to generate sub-threshold stimulation related potentials (SRPs) and study their complexity[40].

4.3. 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, 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, for example, 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); Zenil et al. (2014); Soler-Toscano et al. (2014); Zenil et al. (2015a)). Power laws, 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[41]) 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.

4.4. 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 com-

[40] In such "interleaved" experiments, the idea is to stimulate the brain using tCS during some period of time (e.g., 10 seconds), and measure EEG responses during the next 10 seconds (Castellano et al., 2017). This process can be repeated, e.g., 100 times, as with other event related potential EEG methods. The resulting data (during interleaved no-stimulation periods to avoid artifacts) can be analyzed using LZW compression just as Casali et al. (2013) did, or in other related manners.

[41] Indeed, at the micro-scale level it is believed that groups of ∼100 neurons are organized into mini-columns, and groups of ∼ 75 mini-columns are organized into columns, with evidence from both cytoarchitecture and MRI of modular structure at larger scales (see Friedman and Landsberg (2013) and references therein).

plexity. 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).

5. Discussion

KT proposes a mathematical framework[42] 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[43]. 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[44]. Self-awareness is seen to arise naturally

[42] As already discussed, KT relies on a mathematical substrate rather than a physical one. This seems rather consistent with the possibility of creating conscious AI from essentially algorithms (mathematics).

[43] Although the ultimate physical laws may be simple, also the effective, "coarse grained" laws that derive from them need to be simple to be used in practice by agents. Perhaps it is a logical necessity that such effective physical theories (those that are approximations of the ruling equations at larger scales) are also simple (see, e.g., Israeli and Goldenfeld (2004)), but this not obvious. An interesting challenge is to show that simple, hierarchical rules become again hierarchical and simple after coarse graining. Seems like a reasonable conjecture from the point of view of "conservation of complexity" (complexity can only decrease after coarse-graining). On the other hand, the fact that agents are limited in resources (pressured by the environment) probably makes the discovery of such effective theories/models an important element in KT, and one that we haven't yet discussed.
Ultimately, in KT structured experience is seen to stem from the world, and may be limited by it. If the world were a random number generator (binary digits of π) written in a few lines of code, our structured experience of it would be quite limited, the simplifying model being too simple. Of course, it may well be that brains can become capable of constructing models beyond those describing our observable reality (e.g., the mathematics of complex numbers).

[44] Of course, KT does not explain away the hard problem of consciousness. The theory assumes consciousness is an essential part of reality, and that some systems condense it in a special way. This view is actually fairly in line with work relating to IIT (Tegmark, 2014a; Tononi et al., 2016). Let us summarize the questions KT attempts to answer:

1. When does structured experience take place? *When a compressive model is found and selected.*
2. What are we conscious of? *Models.*
3. Why are experiences unique? *Models are complex objects and each is quite unique.*
4. Why can they be seen as irreducible? *Models are simple programs and cannot be broken into parts.*

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[45]. 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 $\mathcal{K}$) as well as the mutual algorithmic information of world data with present or past external brain inputs and outputs constitute key elements or the development of metrics of consciousness[46]. However, more

5. Why do some experiences feel more real than others? *Experiences feel more real if agents have access to models that fit the data well in a simple and comprehensive way.*
6. Why are conscious of some things rather than others? *Some models are more compressive than others and hence "chosen".*
7. Is self-modeling inevitable? *An agent interacting with the world will perform better (and thus have greater chances for survival) using a self-model, so self-modeling and hence self-awareness are probable byproducts of natural selection.*
8. Why is structured consciousness useful? *Structured consciousness is equivalent to successful modeling in a complex looking world. If we accept that good models feel more real when run, "feeling real" and being more relevant for survival become equivalent.*
9. What types of systems may be capable of structured experience? *Agents as defined above.*
10. What do agents model? *That which they can use to enhance their chances of survival/reproduction.*
11. What are the NCCs? The neural correlates of consciousness (NCC) constitute the minimal set of neuronal events and mechanisms sufficient for a specific conscious percept. With regard to mechanisms, the theory implies that conscious agents must have access to data, instantiate computers that find then run compressive programs modeling world-generated data, compare their outputs with data, and make decisions and learn. These elements lead, for example to RNNs as a valid (universal) paradigm. Neural networks, for example, provide Bayesian estimates of class membership (Richard and Lippmann, 1991; Saerens et al., 2002), so they implicitly carry out this data comparison matching process. Physically, the NCCs of consciousness are thus RNNs that implement simple models and contrast them with data. The last element may be the objective function in the Action Module, which may also be a key NCC, providing valence to experience. KT should eventually provide clues on how models are created by agents, and how indeed, agents may arise from information chaos and in which types of universes that may happen. E.g., in a universal CA which has been randomly initialized, how do agents subjected to natural selection and replication arise?

[45] Such data strings are extremely rare in a probabilistic sense. "The expected value of the Kolmogorov complexity of a random sequence is close to the Shannon entropy" (Cover and Thomas (2006)).

[46] An important practical implementation hurdle for this program is the fundamental uncomputabil-

precise statements should be possible: the connections between algorithmic complexity and recursion with other complexity measures (e.g., power-laws, small-world network behavior, fractal dimensions, etc.) 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 neural networks) 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[47]. 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.

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 110 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

Albantakis, L. and Tononi, G. (2015). The intrinsic cause-effect power of discrete dynamical systems—from elementary cellular automata to adapting animals. *Entropy*, 17(54725502).

Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(47).

ity of $\mathcal{K}$. However, limited framework for computation (such as primitive recursive functions (Ruffini, 2016) can provide upper bounds for this quantity as, e.g., derived from LZW, and thus the means for computable metrics. Undoubtedly, future work will produce more powerful techniques and better bounds for $\mathcal{K}$ from, e.g., machine learning techniques such as deep compressive autoencoders or evolutionary programming.

[47]The principles proposed here to study and eventually quantify human consciousness could be used in other species, machines or to detect signs of life from extraterrestrial signals. For example, electromagnetic emissions from intelligent species in exoplanets should differ from random natural noise in terms of AIT derived metrics.

Andrillon, T., Poulsen, A. T., Hansen, L. K., Leger, D., and Kouider, S. (2016). Neural markers of responsiveness to the environment in human sleep. *The Journal of Neuroscience*, 36(24):6583– 6596.

Atienza, M., Cantero, J. L., Grau, C., Gomez, C., Dominguez-Marin, E., and Escera, C. (2003). Effects of temporal encoding on auditory object formation: a mismatch negativity study. *Cognitive Brain Research*, 16:359–371.

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

Baars, B. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Baker, D. H. and Graf, E. W. (2009). Natural images dominate in binocular rivalry. *PNAS*, 106(13).

Barbour, J. (1999). *The end of time*. Oxford University Press.

Bekinschtein, T. A. and et al., S. D. (2009). Neural signatures of the conscious processing of auditory regularities. *PNAS*, 106(5):1672–1677.

Benidixen, A. and Schröger, E. (2008). Memory trace formation for abstract auditory features and its consequences in different attentional contexts. *Biological Psychology*, 78:231–241.

Bennet, C. (1988). Logical depth and physical complexity. In Herken, R., editor, *The Universal Turing Machine–a Half-Century Survey*, pages 227–257. Oxford University Press.

Berto, F. and Tagliabue, J. (2012). Cellular automata. In (ed.), E. N. Z., editor, *The Stanford Encyclopedia of Philosophy*. Stanford, http://plato.stanford.edu/archives/sum2012/entries/cellular-automata.

Blake, R. and Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, 3.

Blake, R. and Tong, F. (2008). Binocular rivalry. *Scholarpedia*, 3(12):1578.

Blanke, O., Mohr, C., Michel, C. M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T., and Thu, G. (2005). Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. *The Journal of Neuroscience*, 25(3):550–557.

Blanke, O., Ortigue, S., Landis, T., and Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, 419.

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

Bushara, K. O., Grafman, J., and Hallett, M. (2001). Neural correlates of auditory–visual stimulus onset asynchrony detection. *The Journal of Neuroscience*, 21(1):300–304.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., and Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med*, 5(198).

Castellano, M., Ibanez-Soria, D., Kroupi, E., Soria-Frisch, A., Dunne, S., Valls-Sole, J., Verma, A., and Ruffini, G. (2017). Influence of burst tACS on the neural oscillations and detection of change in visual task. In *Brainstim 2017*.

Chaitin, G. J. (1991). Perspectives in biological complexity. In Solbrig, O. T. and Nicolidis, G., editors, *Perspectives in biological complexity*, chapter Algorithmic information & evolution, pages 51–60. IUBS Press.

Chaitin, G. J. (1995). Randomness in arithmetic and the decline and fall of reductionism in pure mathematics. In Cornwell, J., editor, *Nature's imagination*, pages 27–44. Oxford U. Press.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219.

Chaudhary, U., Xia, B., Silvoni, S., Cohen, L. G., and Birbaumer, N. (2017). Brain–computer interface–based communication in the completely locked-in state. *PLOS Biology*, 15(1).

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204.

Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems*, 15:1–40.

Copeland, J. B. (2015). The Church-Turing Thesis. *The Stanford Encyclopedia of Philosophy*.

Costa, M., Goldberger, A. L., and Peng, C.-K. (2002). Multiscale entropy analysis of complex physiologic time series. *Phys. Rev. Lett.*, 89(6).

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & sons, 2nd edition.

Crick, F. C. and Koch, C. (2005). What is the function of the claustrum? *Phil. Trans. Roy. Soc. Lond*, B(380):1271–9.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4):303–314.

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

Deutsch, D. and Marletto, C. (2014). Constructor theory of information. *Proceedings of the Royal Society A*, 471(2174).

Dieter, K. C. and Tadin, M. D. M. D. (2016). Perceptual training profoundly alters binocular rivalry through both sensory and attentional enhancements. *PNAS*, 113(45).

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A. C. A., Baliki, M., and Apkarian, A. V. (2005). Scale-free brain functional networks. *Physical Review Letters*.

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T., Galanaud, D., Puybasset, L., Bolgert, F., ad L. Cohen, C. S., Dehaene, S., and Naccache, L. (2011). Probing consciousness with event-related potentials in the vegetative state. *Neurology*, 77:264–268.

Faugeras, F., Rohauta, B., Weiss, N., Bekinschtein, T., Galanaudd, D., Puybasset, L., Bolgert, F., Sergente, C., Cohen, L., Dehaene, S., and Naccache, L. (2012). Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. *Neuropsychologia*, 50:403–418.

Fekete, T., van Leeuwen, C., and Edelman, S. (2016). System, subsystem, hive: Boundary problems in computational theories of consciousness. *Front. Psychol.*

Fischer, D. B., MD, Boes, A. D., Demertzi, A., Evrard, H. C., Laureys, S., Edlow, B. L., Liu, H., Saper, C. B., Pascual-Leone, A., Fox, M. D., and Geerling, J. C. (2016). A human brain network derived from coma-causing brainstem lesions. *Neurology*, 87.

Fredkin, E. (2003). An introduction to digital philosophy. *International Journal of Theoretical Physics*, 42(2).

Fredkin, E. (2004). Five big questions with pretty simple answers. *IBM J. Res. & Dev.*, 48(1).

Friedman, E. J. and Landsberg, A. S. (2013). Hierarchical networks, power laws, and neuronal avalanches. *Chaos*, 23(013135).

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Science*, 13(7):293–301.

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

Gallos, L. K., Makse, H. A., and Sigman, M. (2012). A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *PNAS*, 109(8):2825–2830.

Gardner, M. (1970). Mathematical games–the fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223:120–123.

Gell-mann, M. and Lloyd, S. (2003). Effective complexity. SFI Working paper 2003-12-068, Santa Fe Institute.

Godwin, D., Barry, R. L., and Marois, R. (2016). Breakdown of the brain's functional network modularity with awareness. *PNAS*, 112(12):3799–3804.

Grau, C., Escera, C., Yago, E., and Polo, M. (1998). Mismatch negativity and auditory sensory memory evaluation: a new faster paradigm. *NeuroReport*, 9:2451–2456.

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

He, B. J. (2014). Scale-free brain activity: past, present, and future. *Trends in Cognitive Sciences*.

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

Hofstadter, D. R. (2007). *I Am a Strange Loop*. Basic Books.

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

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257.

Hudetz, A. G., Liu, X., Pillay, S., Boly, M., and Tononi, G. (2016). Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neurosci Lett.*, 628:132–135.

Husain, M. (2008). Hemineglect. *Scholarpedia*, 3(2):3681.

Hutter, M. (2007). Algorithmic information theory. *Scholarpedia, 2:3 (2007) page 2519, arXiv:cs/0703024*.

Israeli, N. and Goldenfeld, N. (2004). Computational irreducibility and the predictability of complex physical systems. *Phys. Rev. Lett.*, 7(92).

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620–630.

Jaynes, E. (2003). *Probability theory - the logic of science*. Cambridge.

Kanoh, S., Futami, R., and Hoshimiya, N. (2004). Sequential grouping of tone sequence as reflected by the mismatch negativity. *Biol. Cybern.*, 91:388–395.

Kaspar, F. and Schuster, H. G. (1987). Easily calculable measure for the complexity of spatiotemporal patterns. *Phys. Rev. A*, 36(2):842–848.

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

Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, 17:207–321.

Koch, C. and Tononi, G. (2008). Can machines be conscious? *Spectrum*, 45(6):55–59.

Koubeissi, M. Z. and Bartolomei, F. (2014). Electrical stimulation of a small brain area reversibly disrupts consciousness. *Epilepsy & Behavior*, 37:32–35.

Kovács, I., Papathomas, T., Yang, M., and Fehér, A. (1996). When the brain changes its mind: Interocular grouping during binocular rivalry. *PNAS*, 93(26).

Lagercrantz, H. and Changeux, J.-P. (2010). Basic consciousness of the newborn. *Seminars in perinatology*, 34(3):201–206.

Laureys, S. (2005). The neural correlate of (un)awareness: lessons from the vegetative state. *TRENDS in Cognitive Sciences*, 9(12).

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, IT-22(1):75–81.

Li, M. and Vitanyi, P. (1997). *An introduction to Kolmogorov Complexity and its applications*. Springer.

Li, M. and Vitanyi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*. Spriger Verlag.

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

Lloyd, S. (2002). The computational capacity of the universe. *Phys.Rev.Lett.*, 88.

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

Manor, B., Costa, M. D., Hu, K., Newton, E., Starobinets, O., Kang, H. G., Peng, C. K., Novak, V., and Lipsitz, L. A. (2010). Physiological complexity and system adaptability: evidence from postural control dynamics of older adults. *J Appl Physiol*, 109:1786–1791.

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., and Tononi, G. (2005). Breakdown of cortical effective connectivity during sleep. *Science*, 309.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518.

MV Lambert, M Sierra, M. P. and David, A. (2002). The spectrum of organic depersonalization: a review plus four new cases. *The Journal of neuropsychiatry and clinical neurosciences*, 14(2):141–54.

Näätänen, R. and et al. (1978). Early selective-attention on evoked potential reinterpreted. *Acta Psychol Amst*, 42(313–329).

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

Nishimoto, S., Vu, A. T., Naselaris, T., Benjaminia, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21:1641–1646.

Pusiol, G., Esteva, A., Hall, S. S., Frank, M. C., Milstein, A., and Fei-Fei, L. (2016). Vision-based classification of developmental disorders using eye-movements. In *MICCAI2016*.

Quiroga, Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Ravasz, E. and Barabasi, A.-L. (2003). Hierarchical organization in complex networks. *Phys. Rev.*, E(67).

Ray, C., Ruffini, G., Marco-Pallarés, J., Fuentemilla, L., and Grau, C. (2007). Complex networks in brain electrical activity. *European Phys. Lett*, 79(38004).

Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, 44:112–131.

Rendell, P. (2014). *Turing machine universality of the game of life*. PhD thesis, University of the West of England.

Richard, M. D. and Lippmann, R. P. (1991). Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, 3(461-483).

Rovelli, C. (2015). Relative information at the foundation of physics. In Aguirre, A., Foster, B., and Merali, Z., editors, *It from Bit or Bit from It? On Physics and information*, pages 79–86. Springer.

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

Ruffini, G. (2009). Reality as simplicity. *arXiv: 0903.1193*.

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

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Any reasonable cost function can be used for a posteriori probability approximation. *IEEE Transactions on Neural Networks*, 13(5):1204–1210.

Sanchez-Vives, M. V. and Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6.

Schartner, M., Seth, A., Noirhomme, Q., Boly, M., Bruno, M.-A., Laureys, S., and et al. (2015). Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia. *PLoS ONE*, 10(8).

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K., and Muthukumaraswamy, S. D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific Reports*, 7(46421).

Seager, W. and Allen-Hermanson, S. (2015). Panpsychism. *The Stanford Encyclopedia of Philosophy*.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Science*, 17(11):565–573.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2):97–118.

Seth, A. K. (2016). The real problem. *Aeon.co*.

Shic, F., Chawarska, K., Bradshaw, J., and Scassellati, B. (2008). Autism, eye-tracking, entropy. In *7th IEEE International Conference on Development and Learning, 2008. ICDL 2008.*

Siegelmann, H. T. and Sontag, E. (1995). On the computational power of neural nets. *Journal of computer and system sciences*, 50(1):132–150.

Sierra, M., Lopera, F., Lambert, M. V., Phillips, M. L., and David, A. S. (2002). Separating depersonalisation and derealisation: the relevance of the "lesion method". *J Neurol Neurosurg Psychiatry*, 72:530–532.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philos Trans R Soc Lond B Biol Sci.*, 364(1535):3549–3557.

Soler-Toscano, F., Zenil, H., Delahaye, J.-P., and Gauvrit, N. (2014). Calculating Kolmogorov complexity from the output frequency distributions of small Turing machines. *PLoS ONE*, 9(5).

Stanley, G. B., Li, F. F., and Dan, Y. (1999). Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience*, 19(18):8036–8042.

Stiefel, K. M., Merrifield, A., and Holcombe, A. O. (2014). The claustrum's proposed role in consciousness is supported by the effect and target localization of Salvia divinorum. *Front Integr Neurosci.*, 8(20).

't Hooft, G. (2014). *The Cellular Automaton Interpretation of Quantum Mechanics*, volume http://arxiv.org/pdf/1405.1548v3.pdf. arXiv.

Taylor, P., Hobbs, J. N., Burroni, J., and Siegelmann, H. T. (2015). The global landscape of cognition: hierarchical aggregation as an organizational principle of human cortical networks and functions. *Scientific Reports*, 5(18112).

Tegmark, M. (2014a). Consciousness as a state of matter. *arXiv:1401.1219v2*.

Tegmark, M. (2014b). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Knopf.

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17:450–461.

Tononi, G. and Koch, C. (2008). The neural correlates of concsciousness - an update. *Ann. N.Y. Acad. Sci.*, 1124:239–261.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(42):230–265.

Uhrig, L., Dehaene, S., and Jarraya, B. (2014). A hierarchy of responses to auditory regularities in the macaque brain. *The Journal of Neuroscience*, 34(4):1127–1132.

van den Heuvel, M. P. and Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12):683.

Van Gulick, R. (2016). Consciousness. *The Stanford Encyclopedia of Philosophy*, Winter 2016.

Vitányi, P. M. B. and Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov Complexity. *IEEE Transactions on Information Theory*, 46(2):446–464.

Wallace, C. S. and Dowe, D. L. (1999). Minimum message length and Kolmogorov Complexity. *The Computer Journal*, 42(4).

West, G. B. (1999). The origin of universal scaling laws in biology. *Physica A*, 263:104–113.

Wheeler, J. A. (1990). Complexity, entropy, and the physics of information. In Zurek, W. H., editor, *Complexity, Entropy, and the Physics of Information*, chapter "Information, physics, quantum: The search for links". Addison-Wesley.

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.

Wolpert, D. (2008). The physical limits of inference. *Physica D: Nonlinear Phenomena*, 237(9):1257–1281.

Yu, K. and Blake, R. (1992). Do recognizable figures enjoy an advantage in binocular rivalry? *J Exp Psychol Hum Percept Perform. 1992 Nov;18(4):1158-73.*, 18(4):1158–73.

Zenil, H., Kiani, N. A., and Tegnér, J. (2015a). Methods of information theory and algorithmic complexity for network biology. *arXiv:1401.3604v7*.

Zenil, H., Marshall, J. A., and Tegnér, J. (2015b). Approximations of algorithmic and structural complexity validate cognitive-behavioural experimental results. *arXiv:1509.06338*.

Zenil, H., Soler-Toscano, F., Dingle, K., and Louis, A. A. (2014). Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks. *arXiv:1306.0322v3*.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences by variable rate coding. *IEEE Transactions on Information Theory*, IT-24:530–536.

Zuse, K. (1967). Rechnender raum. *Elektronische Datenverarbeitung*, 8:336–344.