Towards an algorithmic information theory of consciousness (KT)

Giulio Ruffini, Edmundo Lopez-Sola, Roser Sanchez-Todo

Brain Modeling Department
Neuroelectrics Barcelona, Barcelona

giulio.ruffini@neuroelectrics.com

April 19, 2022

Overview

Motivation

The subjective route to KT: Experience (1P)

We start from the fact of *experience*, from the first person, subjective standpoint (Ruffini, 2017).

Meditation, psychedelics, religious experiences (e.g., Buddhism) suggest that experience can be pure/primordial, free from mental constructs such as the ego.

From the self-evidence of our own experience, the "what it's like to be", we assume that primordial experience does not allow for or require prior causes.

Warning: We assume *there exists experience.*

KT does not address the hard problem of consciousness.

Structured experience ($\mathcal{S}$)

We aim to build a theory around the notion of *structured experience*—where *mathematics* and experience meet.

Mathematics: The science of structure, order, and relation that has evolved from counting, measuring, and describing the shapes of objects (Gray (2010), Enc. Britannica).

We observe that experience is *structured*: at least during wakefulness, there is a spatial, temporal, and conceptual organization of our first-person experience of the world and of ourselves.

Definition (*structured experience* ($\mathcal{S}$))

The phenomenal structure of consciousness encompassing both sensory qualia and the spatial, temporal, and conceptual organization of our experience (Van Gulick, 2016).

Scientific strategy for the study of $\mathcal{S}$

This definition of $\mathcal{S}$ can be empirically explored in reporting humans, and we aim to characterize it with methods generally applicable to a broad range of systems.

The general strategy will be to quantify the structure of experience from **first person reports** (1P) in humans and attempt to associate it with 3P state measurements (e.g., EEG or fMRI) or behavior using mechanistic insights derived from neuroscience and mathematics.

With this knowledge at hand, we can then study $\mathcal{S}$ in other systems (non-reporting humans, other living species or artificial agents), and provide an educated guess about the agent's actual experience.

The objective route (3P): persistence and life

We can also start by attempting to define what *life* is.

What remains after the passage of eons must rightfully be called a *persistent pattern*.

There may be several types of such patterns (= our models). Some seem static and impervious to the world, such as protons.

Definition (Life)

Patterns that readily interact with the world but persist by partly capturing structure in the world they inhabit to maintain or replicate (homeo- and meta-homeostasis).

The connection with the first viewpoint is that, in KT, this generalized *life is what is capable of* $\mathcal{S}$.

As part of our framework, we should study the mathematics of the emergence of life.

AIT and Kolmogorov complexity

Kolmogorov complexity ($\mathcal{K}$) I

We can think of agents as physical systems, and in turn, of physical systems as dynamical systems implementing functions. This allows us to analyze agents from the standpoint of computation theory.

Warning: Computation is a mathematical concept (Turing machine)

The use of computational framework in KT should not be construed to mean that the brain is literally a physical von Neumann computer (such as a laptop).

A computational perspective leads us directly into algorithmic information theory (AIT) and its central concept:

Definition (Kolmogorov complexity of a dataset $\mathcal{K}$)

The length of the shortest program capable of generating the dataset (Kolmogorov, 1965; Cover and Thomas, 2006).

Kolmogorov complexity ($\mathcal{K}$) II

Data

['3.14159265358979323812644133278315385134669383751090189572 6',
'0183844990527447207158188099987323024033062286954749281803 20',
'7536741059927308891818990028396031781938117734317227203959 358',
'2681413334524778875085102698279378964782750493870397771769 00',
'4052415221610820855312843745938676064382634796104288333732 40',
'8145257445815990368592973580473662870200472728373853305809 90',
'3219954525817261060491726903951980625736652628712969609052 98',
'7671009626536868727705897236681911022946747224333945197705 300',
'0551194650093595826117491241484296185084399045107683259827 42',
'2796752757654895253426715085241676350146131019255952023557 55',
'1025857702005465094718990466067932265266791062245718982808 00',
'1079018037270642344517933906555987110685595058411279280600 32',
'3371928865285029666491930322116158736056828346647017547248 799',
'5047881060144389117952834809580569970793136663404028823381 00',
'8859878642076502910551119205187427530885652009759019838144 337',
'9959543614228588756397294345051927549776778582253975174629 84',
'1062524121123966056730332965125414383488745839700585848528 01',
'7589679472248565381117609951275291347605491550540934846424 04',
'1446507660740795192052525066129611766269075058004912488392 16',
'0959876470338606019660536434366214015643677347934323381160 39',
'0976580751658357898726311389441335589100941109793965014816 82',
'5098139534461506842152558590766974227268103727888582250211 20',
'0915431100227729084163271858880408201234196220085730263522 49',
'6617415036846819539299612901084677056805139712488551953236 0',
'4226226703436236014338139502905372278408088211001260582911 94',
'0710443484622851371228598214309735550913149138931732985440 09',
'7640031876246093354087677648166725209463342004957077468323059 9',
'4035067204193117534783667673388479448653806180626862653194 97',
'3434402584778120303661390088673738248660363539288600905926 86',
'3144203920327257440556024223421528067656224653298329717620 922',
'0959331949526318173011107900898395940425054357988500745683 20',
'2087643303992130536656797476045645065726401551750656694184 4',
'9484861086981528759507365950709709350354775357654702788250 68',
'6472743448866089944433814954585492799023580238978785646415 88',
'8274199009398310215360665972538336186638618863539243111550 72',
'5465632597341871693092005048369162816989158228242909338067 93',
'4236346495476442304739861284858223431494392904021289244957 74',
'5968725303622465325518055916047926857762066107005081060125 00',
'1379937745532168307261807405369608550927747786054095875584 1',
'7126238760037876516321431973866779929212623501645513508352 38',
'8611813000990276977397495558044340651038911414431714386451 33',
'2369551974083304070317338997181759848124743853743856318798 66',
'9771169973976839394704624349579551136117614953533238626691 238',

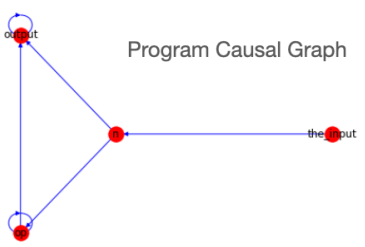
Program/model

```python
# https://www.wikihow.com/Write-a-Python-Program-to-Calculate-Pi
def nilakantha(the_input):

    variables=['the_input', 'n','op','output']
    dictvariables = { i : variables[i] for i in range(0, len(variables) ) }
    c = np.zeros((len(variables),len(variables)),dtype='int32') #from to

    output = Decimal(3.0)
    op = 1
    n = 2

    c[1,0] += 1
    for n in range(2, 2*the_input+1, 2):
            c[3,1 ] += 1; c[3,2 ] += 1; c[3,3] += 1
            output += 4/Decimal(n*(n+1)*(n+2)*op)
            c[2,2] += 1; c[2,1] += 1;
            op *= -1
    return output,c.transpose(),dictvariables
```



Mutual algorithmic information ($\mathcal{M}$)

With $\mathcal{K}$ at hand, we can define an algorithmic version of mutual information we will also need:

Definition (Mutual algorithmic information complexity $\mathcal{M}$)

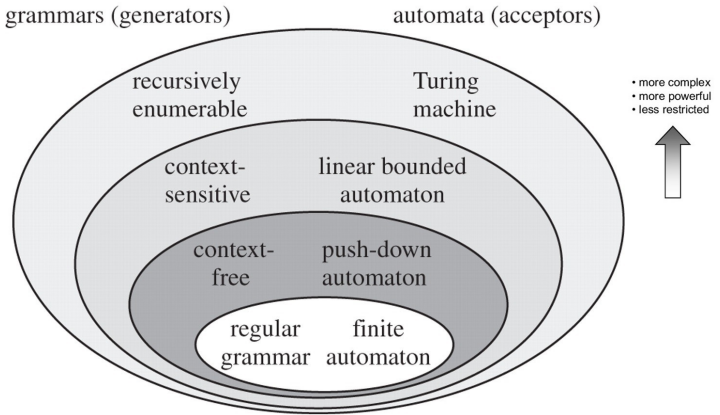
The *mutual algorithmic information* $\mathcal{M}(x:y)$ *between two strings* x *and* y*, is given by*

$$\mathcal{M}(x:y) = \mathcal{K}(x) + \mathcal{K}(y) - \mathcal{K}(x,y),$$

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).

Hierarchy class (Fitch, 2014)

Not all programming languages are equal. Recurrence is needed for Turing completeness, for example.



Model I

The notion of **model** is central in KT and other theories of consciousness.

Definition (Model)

A model of a dataset is any program that generates the dataset.

Models may differ in two ways: they may implement different functions, or they may implement the same function in different ways. Both aspects matter here. We focus on those that implement the right functions succinctly.

Definition (Optimal model)

The optimal model of a dataset is the shortest program that generates (or, equivalently, compresses) the dataset.

Model II

An **optimal model needs to capture and exploit all the structure in the dataset**—and nothing else. In some sense, the *structure of the model can be described by the group of symmetries of the dataset* (Ruffini, 2016).

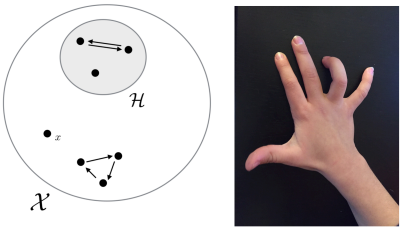
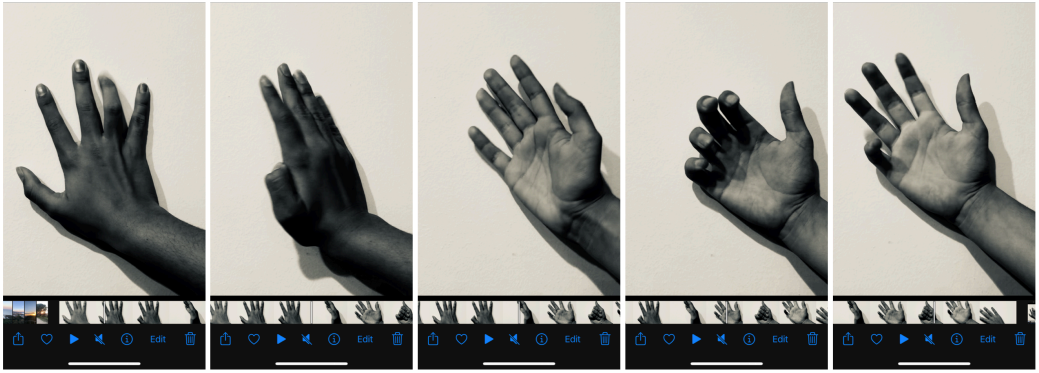


FIGURE 1. Left: Input (image) space $\mathcal{X}$, the subspace of hand images $\mathcal{H} = \{x \in \mathcal{X} | f(x) = 1\}$ and a example of automorphism (arrows) leaving the class one set $\mathcal{H}$ invariant. Right: a sample element $x \in \mathcal{H} \subset \mathcal{X}$.

Suppose we are given a stack of images of a hand, e.g., from the frames of a movie of a moving hand created using a generating function, $y = f(\theta)$, where y is the image in a frame and θ a parametrization of the hand image and view.

Model III



The structure of the dataset is encapsulated by the function f, or the minimal program that encodes it and that constitutes the *invariant* object. The **symmetries** of the dataset are parametrized by θ, and they are symmetries in the sense that if $\theta \to \theta'$ and $y \to y'$, then the equation $y' - f(\theta') = 0$ holds true.

Why are "good models" good?

The rationale for the importance of compressive (succinct) models is discussed in detail in Ruffini (2009, 2017). Rephrase of the principle of Occam's Razor: *one should not increase, beyond what is necessary, the number of entities required to explain anything.*

The universe appears to be simple. Simple rules can create apparent complexity.

Simple data generators are more likely if the universe rules are drawn from a random algorithmic bingo (Solomonoff's prior).

Simple models are less biased and generalize better (Occam, Laplace, Jaynes).

They are easier to construct.

More economical, easier to store, use and reuse for model-building.

The central hypothesis

The central hypothesis in KT

The central hypothesis of KT

The **central hypothesis** in KT is that an agent has $\mathcal{S}$ (i.e., lives stronger, more structured experiences) to the extent it has access to *encompassing and compressive models* (*good models*) to interact with the world.

More specifically, *the event of structured experience arises from the act of successfully comparing good models with data.*

Where does the structure of experience come from? Program characteristics deriving from the hierarchy class it belongs to, its structure and length, determine the properties of structured experience.

The central hypothesis in KT

Compressive model refers to a simple, succinct program—the optimal model defined above is best possible scenario. Compressive models are special: to be short and accurate they need to capture structure in the data.

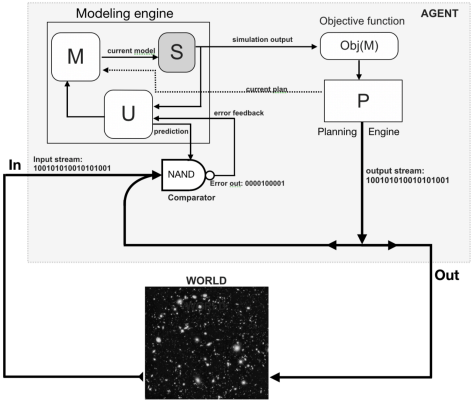
Encompassing refers to the amount of data the agent's model successfully matches information from the world, and hence to its explanatory potential. Eq., high $\mathcal{M}$. Models need to account for all the data available to the agent and do so with little error.

Models are constructed from information generated by the sensorimotor system as the agent interacts with the external world: the agent's model can/should account for data generated by the external world and by the agent itself—i.e., include a *self-model*.

The algorithmic agent

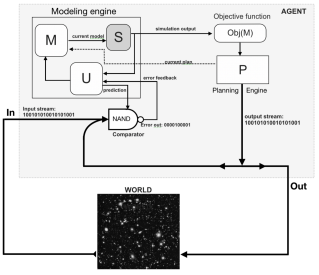
The algorithmic agent

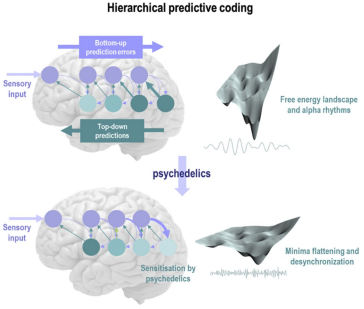
Minimal set of elements needed for a homeostatic algorithmic system in an information bath. Can be connected with neurobiology.

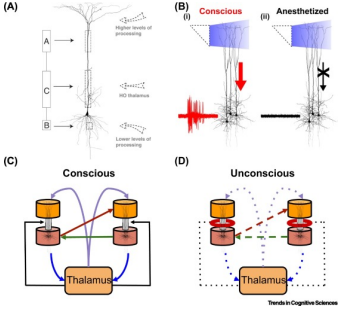


The algorithmic event of $\mathcal{S}$

Theories of cortical processing emphasize the separation of forward and backward information flow in the cortex also mirrored at the level of single cortical pyramidal cells (Carhart-Harris and Friston, 2019; Aru et al., 2020). **Comparator** implemented hierarchically in L5 P cells.







Model-building: life

How do agents build models? In addressing this, we are led to connecting the concepts of life, intelligence, and $\mathcal{S}$.

Both life and intelligence represent processes to construct simple models for the persistence of algorithmic information-preserving systems across time.

Starting from resilient building blocks (static persistence), from a computational perspective *life* is an algorithmic process: program building carried not solely by the individual agent, but by the transgenerational agent through evolution for *meta-homeostasis* (preservation of kind) (v. also Walker and Davies (2013); Chaitin (2012)).

Model-building: intelligence

Evolutive pressure gives rise to the next leap, *intelligence*: agents that, starting from their static model (DNA in life) build higher-level compressive models of the world within their lifetime, e.g., using brains.

Importantly, KT holds that both static-model and active-modeling agents enjoy structured experience, only that their level of structure is possibly different.

[What comes after life and intelligence?]

Model-building in artificial agents

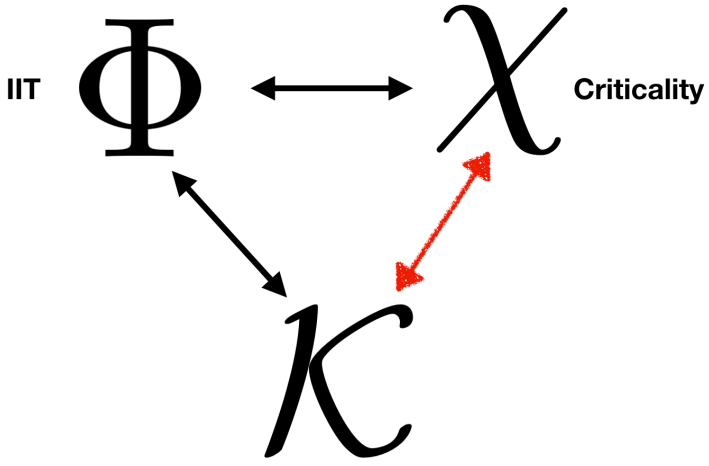
As a consequence of the above, we should explore two routes to the construction of artificial model-building agents:

A) **Single generation** model building where agents are endowed with a *simplicity bias* (this is what we call the *intelligence* approach).

B) **Transgenerational** model building (*life*) where the bias for simplicity is not added by hand but emerges naturally from the construction process that favors simple short programs under evolutionary pressure in *environments governed by simple laws*.

Computation and dynamics

The criticality view I



The criticality view II

Computation in Nature is carried out by dynamical systems with very large degrees of freedom.

Brains operate close to such critical boundaries consistent with the notion of self-organized criticality (SOC) (Bak et al., 1988; Chialvo, 2004; Cocchi et al., 2017; Carhart-Harris, 2018; Deco et al., 2021).

Altered state of consciousness appear to move the brain away or towards criticality (Carhart-Harris and Friston, 2019; Ruffini and et al, 2022).

That KT and criticality theory may be linked arises from the observation that computational structures—dynamical systems—instantiating simple, compressive models of compositional data that exhibits regularities/symmetries (conserved quantities) must have special properties. What are they?

The criticality view III

Recall the movie of a moving hand in empty space, $y(t) = f(\theta(t))$.

Although y may be embedded in a very high dimensional space, its dimension is actually very small if the set of parameters θ controlling the hand function is small.

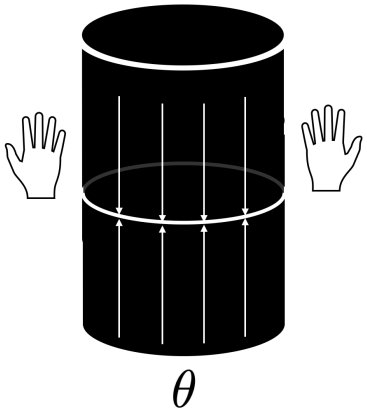
The state of a dynamical system generating frames of the moving hand, regardless of how large its natural space is (e.g., involving a large number of neurons) must also lie in a low dimensional subspace, a reduced manifold.

Near criticality ($\mathrm{Re}[\lambda] \sim 0$) the dynamics of complex systems collapse to low dimensional manifolds. (Jirsa, 2020; Jirsa and Sheheitli, 2022).

In a Hamiltonian dynamical system where $g = y(t) - f(\theta(t)) = 0$ (the constraint), Noether's theorem states that H is invariant under the group of transformations generated by g (symmetry) (Dirac, 2001; Jose and Saletan, 1998)).

The criticality view IV—the center manifold

A hand is a hand is a hand.



The criticality view V

Structure in data, the collapse of dynamics to low-dimensional spaces, criticality and associated features such as maximal information flow, power laws, long time scales and enhanced susceptibility, $\mathcal{K}$ and $\mathcal{S}$ are thus deeply connected.

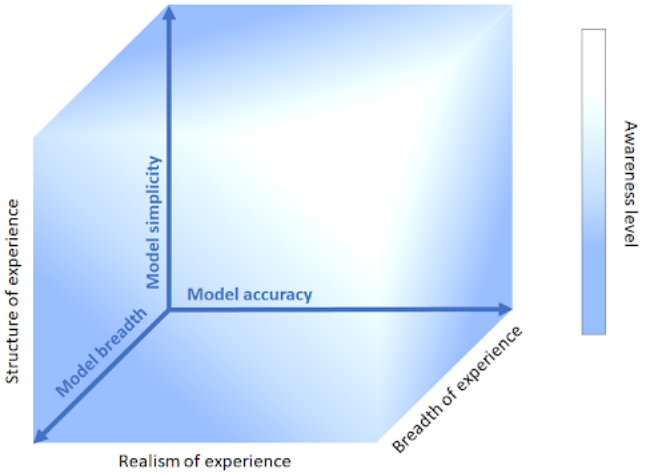
The dimensionality and manifold structure of the reduced dynamics together with the mutual information of the system with the external world provide, respectively, metrics on the simplicity of the models and the amount of algorithmic information captured.

The structure of the reduced manifold maps into the structure of experience, while model accuracy and breath ($\mathcal{M}$) map into the realism and breath of experience. These are the three dimensions of $\mathcal{S}$.

The agent-world lock loop, tracking real world (structured) data, helps to keep dynamics on the reduced manifold ($\mathcal{S}$!). Psychedelics, meditation, sensory deprivation may "lift" dynamics up from enslaved dynamics.

Neurophenomenology of KT

The dimensions of structured experience



Altered states

Neurophenomenology defines a methodological strategy for integrating phenomenological and neurobiological accounts under one research program, by bridging two irreducible phenomenal domains: 1P (phenomenology, subjective measures) and 3P (physiology, objective measures) data (Varela, 1996).

Altered states of consciousness such as psychedelics or meditation offer an interesting context to study the effect of perturbing the mechanisms of $\mathcal{S}$.

Meditative states are associated with the global dissolution of the embodied self (Millière et al., 2018) and can serve as unique models for a neurophenomenological investigation of self-dissolution (disengagement of self-models).

As an objective measure of structured experience, we can analyze descriptive narratives in speech form through state-of-the-art computational analysis (NLP) to establish metrics on text structure such as semantic coherence and speech disorganization index (Sanz et al., 2021; Tagliazucchi, 2022; Mota et al., 2017).

Closing

Philosophy

KT is most naturally viewed in the context of panpsychism ('mind is everywhere', see Strawson, Goff (2019)), a particular, somewhat controversial version of the philosophy of consciousness.

Idealism is another, perhaps more rigorous philosophical background (consciousness as the fundamental entity, 'mind is everything').

Although not necessary to explore the scientific implications of the theory, the adoption of these can be motivated by simplicity, consistency criteria (Symes, 2022).

Ethics

KT does not give any special place to humans. All systems that capture structure from the world have structured experience. Pleasure/pain associated with the **objective function** O.

Morality: we have natural notions of *good* or *evil* in computational terms based on the behavior of objective functions.

E.g., , we may say that Agent's A is *evil or morally wrong* to Agent B if the objective function of A, O_A, increases when O_B decreases, that is $O_A(O_B)$ is decreasing or $O'_A(O_B) < 0$.

Conversely, we say that Agent A is *good or morally right* to Agent B if $O'_A(O_B) > 0$.

Synergistic behavior emerges when agents are good to each other, and mutually-destructive behavior takes place in the complementary case.

Future

Much work remains to be done! Will KT provide a unification framework for the different approaches to consciousness? IIT, GWT, FEP, DFT all seem to fit.

Can we evolve agents in computational frameworks? Are persistent patterns unavoidable (KT conjecture) in computation? Are there types other than static (proton), life and intelligence?

Can we further develop a theory bridging dynamical systems and AIT?

Can we discover the structure of reduced dynamics from physiological (system state) data?

Can we design neurophenomenological methods to discover the structure of reports and behavior?

Can we design model-building agents mimicking life or intelligence? Is AI the next evolutionary model-building leap

Thanks

Thanks for your attention and curiosity!

giulio.ruffini@neuroelectrics.com, @ruffini (Twitter)

Slides available at
https://github.com/giulioruffini/Ruffini-KT-Tucson-presentation-April-19-2022

References I

Aru, J., Suzuki, M., and Larkum, M. E. (2020). Cellular mechanisms of conscious processing. *Trends in Cognitive Sciences*, 24(10):814–825.

Bak, P., Tang, C., and Wiesenfeld, K. (1988). Self-organized criticality. *Physical Review A*, 38(1):364–374.

Carhart-Harris, R. L. (2018). The entropic brain - revisited. *Neuropharmacology*.

Carhart-Harris, R. L. and Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews*, 71(3):316–344.

Chaitin, G. J. (2012). *Proving Darwin*. Pantheon, New York, NY.

Chialvo, D. R. (2004). Critical brain networks. *Physica A*, 340:756–765.

Cocchi, L., L.Gollo, L., Zalesky, A., and Breakspearac, M. (2017). Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in Neurobiology*.

References II

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

Deco, G., Perl, Y. S., Sitt, J. D., Tagliazucchi, E., and Kringelbach, M. L. (2021). Deep learning the arrow of time in brain activity: characterising brain-environment behavioural interactions in health and disease.

Dirac, P. (2001). *Lectures on Quantum Mechanics*. Dover Books on Physics. Dover Publications, Mineola, NY.

Fitch, W. T. (2014). Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. *Physics of Life Reviews*, 11(3):329–364.

Goff, P. (2019). *Galileo's Error — Foundations for a New Science of Consciousness*. Penguin Random House UK.

Gray, J. J. (2010). Mathematics.

References III

Grunwald, P. and Vitanyi, P. (2004). Shannon information and kolmogorov complexity. *arXiv:cs/0410002*.

Jirsa, V. (2020). Structured flows on manifolds as guiding concepts in brain science. In *Selbstorganisation – ein Paradigma für die Humanwissenschaften*, pages 89–102. Springer Fachmedien Wiesbaden.

Jirsa, V. and Sheheitli, H. (2022). Entropy, free energy, symmetry and dynamics in the brain. *Journal of Physics: Complexity*, 3(1):015007.

Jose, J. V. and Saletan, E. J. (1998). *Classical dynamics*. Cambridge University Press, Cambridge, England.

Kolmogorov, A. N. (1965). Three approaches to the definition of the concept "quantity of information". *Probl. Peredachi Inf.*, pages 3–11.

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

References IV

Millière, R., Carhart-Harris, R. L., Roseman, L., Trautwein, F.-M., and Berkovich-Ohana, A. (2018). Psychedelics, meditation, and self-consciousness. *Frontiers in Psychology*, 9.

Mota, N. B., Copelli, M., and Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia*, 3(1):1–10.

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).

Ruffini, G. (2017). An algorithmic information theory of consciousness. *Neurosci Conscious*.

Ruffini, G. and et al (2022). Lsd-induced increase of ising temperature and algorithmic complexity of brain dynamics. *BioRXiv*.

References V

Sanz, C., Pallavicini, C., Carrillo, F., Zamberlan, F., Sigman, M., Mota, N., Copelli, M., Ribeiro, S., Nutt, D., Carhart-Harris, R., et al. (2021). The entropic tongue: disorganization of natural language under lsd. *Consciousness and Cognition*, 87:103070.

Symes, J. (2022). *Philosophers on consciousness*. Bloomsbury Academic, London, England.

Tagliazucchi, E. (2022). Language as a window into the altered state of consciousness elicited by psychedelic drugs. *Frontiers in Pharmacology*, page 900.

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

Varela, F. (1996). Neurophenomenology: a methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3(4):330–349.

Walker, S. I. and Davies, P. C. W. (2013). The algorithmic origins of life. *Journal of The Royal Society Interface*, 10(79):20120869.