DOI:
10.1016/j.neubiorev.2019.01.023

Document Version
Peer reviewed version

Manuscript Title:

Conflicting Emergences.
Weak vs. strong emergence for the modelling of brain function.

Authors:

Federico E. Turkheimer, PhD (Institute of Psychiatry, King's College London, UK)

Peter Hellyer, PhD (Institute of Psychiatry, King's College London, UK)

Angie A. Kehagia, PhD (Institute of Psychiatry, King's College London, UK)

Paul Expert, PhD (EPSRC Centre for Mathematics of Precision Healthcare, Imperial College London, UK)

Louis-David Lord (Department of Psychiatry, University of Oxford, UK)

Jakub Vohryzek (Department of Psychiatry, University of Oxford, UK)

Jessica De Faria Dafflon (Institute of Psychiatry, King's College London, UK)

Mick Brammer, PhD (Institute of Psychiatry, King's College London, UK)

Robert Leech, PhD (Institute of Psychiatry, King's College London, UK)

Address for Correspondence:

Professor Federico E. Turkheimer

Institute of Psychiatry, King's College London

Room 3.05, Centre for Neuroimaging Sciences

Institute of Psychiatry, PO89,

De Crespigny Park, London SE5 8AF, U.K.

Telephone / Fax: 020 3228 3051 / 2116

Email: federico.turkheimer@kcl.ac.uk

Sources of Support:

FET is funded by the PET Methodology Program Grant (Ref G1100809/1) and the project grant "Development of quantitative CNS PET imaging probes for the glutamate and GABA systems" from the Medical Research Council UK (MR/K022733/1).

Abstract:

The concept of "emergence" has become commonplace in the modelling of complex systems, both natural and man-made; a functional property" emerges" from a system when it cannot be readily explained by the properties of the system's sub-units. A bewildering array of adaptive and sophisticated behaviours can be observed from large ensembles of elementary agents such as ant colonies, bird flocks or by the interactions of elementary material units such as molecules or weather elements. Ultimately, emergence has been adopted as the ontological support of a number of attempts to model brain function. This manuscript aims to clarify the ontology of emergence and delve into its many facets, particularly into its "strong" and "weak" versions that underpin two different approaches to the modelling of behaviour. The first group of models is here represented by the "free energy" principle of brain function and the "integrated information theory" of consciousness. The second group is instead represented by computational models such as oscillatory networks that use mathematical scalable representations to generate emergent behaviours and are then able to bridge neurobiology with higher mental functions. Drawing on the epistemological literature, we observe that due to their loose mechanistic links with the underlying biology, models based on strong forms of emergence are at risk of metaphysical implausibility. This, in practical terms, translates into the overdetermination that occurs when the proposed model becomes only one of a large set of possible explanations for the observable phenomena. On the other hand, computational models that start from biologically plausible elementary units, hence are weakly emergent, are not limited by ontological faults and, if scalable and able to realistically simulate the hierarchies of brain output, represent a powerful vehicle for future neuroscientific research programmes.

Keywords: brain; emergence; weak emergence; strong emergence; computational models; Bayesian inference; free energy principle; integrated information theory; oscillators; multi-scale.

Prologue.

In scientific inference, complex phenomena arise through interactions among simpler or elementary entities in a process termed “emergence”. In such a process, the properties of the aggregation of the elementary agents that generates the pattern of behaviour are not easily reducible to a combination of the properties of the primitive elements.

Emergence has become a tantalizing topic because many examples of emergent phenomena abound in (but are not limited to) the natural sciences, for example the assembly of complex structures by ant colonies such as bridges and rafts, the swarming behaviours of bees, the flocking behaviour of birds and the murmurations of starlings [Video 1] (Burns et al., 2016; Mlot et al., 2011; Reid et al., 2015). Emergent phenomena in nature can also be seen in weather systems, natural disasters (e.g. typhoons and forest fires), as well as in human-created communities (e.g. cities, the stock market); ultimately the concept of emergence has been offered as a model for human behaviour (Dennett; Miller; West).

Video 1: *Emergent properties of a murmeration of starlings, which follow simple rules in terms of their pair-wise interactions, but together form a complex and adaptive emergent pattern (Video reproduced under licence from Adobe Inc.)*

When studying the brain, we often examine it in a manner that highlights a hierarchy of scales that starts with the cellular milieu (e.g. blood vessels, neurons and glia) with its diverse molecular constituents [Figure 1A]. These building blocks are the elementary components of tissues, nuclei and cortical layers which ultimately are then further arranged into cyto-architectonic regions, often associated in the modern phrenological approach with functional networks [Figure 1B]. Within this hierarchy, each layer, or level of description, exhibits a function that seems autonomous with respect to the activity found at the higher (or perhaps more ‘macroscopic’) level, but that shows a clear dependency on those layers functionally below [Figure 1C]. It follows that the top level of this hierarchy, human behaviour, emerges from interactions within and between these different layers or spatial scales.

Alternatively one can adopt a top-down approach and capture the varieties of perception and action into some general overarching principle that can be assigned to brain tissue and scaled down to the intricacies of receptor systems, metabolism etc.

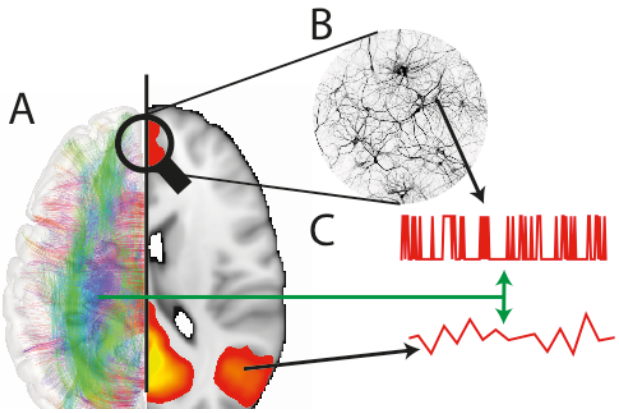

Figure 1: *Our overview of the brain reveals structure and function at multiple spatial and temporal scales **A->B**. Structural connectivity can be explored, both at the macroscopic scale, i.e. regions to region, and at the microscopic level by measuring interactions between cells within and between cortical layers. **A,B -> C**. Each of these macroscopic and microscopic descriptions of function, forms a hierarchy, which reveals different, yet complementary information about the function of the underlying tissue. For example, functional MRI reveals a temporally 'slow' time-course of activity over a wide region of the brain, whereas electrophysiological measures reveal highly detailed spiking time-courses of spatially highly localised tissue. However, these two levels of description are strongly interlinked.*

Irrespective of the approach, emergence has often been used as a conceptual framework to integrate seemingly distant phenomenologies. However, when "emergence" is called into action one can easily fall into logical fallacies that, while extensively debated in the epistemological literature, seem not to be fully recognized in the wealth of current modelling work in the neurosciences and psychology, both in terms of the formulation of traditional 'box and arrow' models of cognition, or in the more recent trend towards the building of large-scale computational simulations of neurobiological function.

Here we review the concept and use of emergence in the experimental neurosciences focusing on two distinct "types" of phenomenological emergence, "strong" and "weak", and their relation to some popular models of brain function: in particular the "free energy principle" (Friston, 2009, 2010; Friston et al., 2006) and "integrated information theory" (Hoel et al., 2013b; Oizumi et al., 2014; Tononi, 2012) as examples of "strong" emergence, and computational oscillatory models as representative of the "weak" emergence (Breakspear, 2017; Deco et al., 2011).

Emergence

Emergence is a contemporary concept with a long history in evolutionary science (for a detailed narrative see Peter Corning 's essay (Corning, 2002)). The concept of emergence was first introduced by the physiologist George H. Lewes in his book *Problems of Life and Mind* (Lewes, 1879, pp. 412) "

… The emergent is unlike its components in so far as these are incommensurable, and it cannot be reduced to their sum or their difference…".

Through this definition, it was possible to form a framework that is generally able to make sense of widely observed leaps in the complexity of nature (Mill, 1874) – particularly in the formation of complex objects from relatively simple elementary parts. For example, it can imaginatively relay how hydrogen and oxygen combined together make one very different molecule of water, convey the punctuated acceleration in taxonomic lineages or be extended to depict the almost unlimited possibilities of a game of chess. However, the 'seeds' of the concept of 'emergence' can be traced much further back in time. Aristotle (Aristotle, 1994) argued that quantitative, incremental changes to the elementary parts of a system or construction may lead to qualitative changes to the whole that are different from, and irreducible to, their parts. The problem is that, by their very nature, such wholes are unpredictable and ultimately their "emergence" is descriptive or richly allusive but fails to explain much if anything about how they come to be. Thus the concept of emergent phenomena rested for many decades on the forgotten shelves of scientific theories in search of a metaphysical foundation, until dynamical system theory produced nonlinear mathematical tools, cellular automata and agent-based models which breathed new life into the idea of modelling interactions within complex systems that were deterministic at the level of interactions among elementary components.

Modern Emergence can be divided into two epistemological types: strong and weak. A system is said to exhibit strong emergence when its behaviour, or the consequence of its behaviour, exceeds the limits of its constituent parts. Thus the resulting behavioural properties of the system are caused by the interaction of the different layers of that system, but they cannot be derived simply by analysing the rules and individual parts that make up the system. Weak emergence on the other hand, differs in the sense that whilst the emergent behaviour of the system is the product of interactions between its various layers, that behaviour is entirely encapsulated by the confines of the system itself, and as such, can be fully explained simply though an analysis of interactions between its elemental units.

The kind of emergence that surfaced first in the neurosciences was greatly shaped by the thinking of Roger W Sperry (1981 Nobel prize in physiology) who proposed a view of the brain characterized by a strong top-down organisational component (Sperry, 1980):

"*…It is the idea, in brief, that conscious phenomena as emergent functional properties of brain processing exert an active control role as causal detents in shaping the flow patterns of cerebral excitation. Once generated from neural events, the higher order mental patterns and programs*

have their own subjective qualities and progress, operate and interact by their own causal laws and principles which are different from and cannot be reduced to those of neurophysiology."

Note that Sperry was adamant that his model did not imply any form of mind brain dualism nor a parallel existence of neurobiological and mental processes but that, after emergence, mental processes would take over and exert control down to the cellular level (Sperry, 1980).

Strong Emergence

The directional dominance of higher versus lower processes makes Sperry's model the paradigmatic exemplar of "Strong Emergence"; a paradigm that comes with epistemological consequences (see collated essays in (Clayton and Davies, 2006)). If the existence of a whole cannot be equated with facts about the distribution and interactions of its particles throughout space and time (along with the laws of physics), then new fundamental laws of nature are needed to explain these phenomena. Indeed in Sperry's model, higher order mental patterns and programs have their own subjective qualities and dynamics and operate by their own laws and principles which are different from and cannot be reduced to those of neurophysiology - they exist and operate in a separate domain, that of psychology (Sperry, 1980).

More recent propositions have followed this path. Predictive processing models (Bubic et al., 2010) argue that the existence of 'expectation states' within a number of cognitive domains of the brain, act in concert with the role of the brain to realise planned events – comparing the subsequent action which as a result of external factors that change expectation may violate the initial prediction of behaviour on a purely feed-forward expectation of cognitive function in the brain. The Bayesian computational model of brain function, also called the "free energy principle" (FEP) (Friston, 2009, 2010; Friston et al., 2006) is an example of such an approach and a paradigmatic exemplar of strong emergence (Lestienne, 2014). In this model, brain-environment interactions of an agent are represented as a loop in which the primary sensory inputs are first processed with prior knowledge of the most probable cause of these signals in a top-down fashion; the brain then combines prior and sensory information and calculates the posterior percept (this process is called Bayesian inversion) that is transmitted to the executive areas of the brain. Within the "executive control system" it is conceived that the feed-back percept is compared against the initial prediction with a gain function (which itself is a realised form of the prior belief set, i.e. learned) that gauges the return of various possible actions onto the environment. This model assumes that in the brain, signals directed from higher to lower levels of the neural hierarchy are more abundant than those directed upwards – a necessary consequence of the postulation that brain activity is dominated by the drive to progressively improve the inferred internal model of cause and effect (i.e. the activity -> behaviour coupling) though the modulation of synaptic connections. From a modelling perspective, the FEP assumes the existence of a number of brain states that parameterize the prior probabilities of the model as well as providing the basis of the gain/loss state function. This largely Bayesian hypothesis formulates perception as a constructive

process based on internal models. As FEP is operated by a set of rules that are treated independently of underlying neurobiology and only loosely constrained (inspired) by metabolic anatomical/neural constraints, FEP can be considered strongly emergent.

Integrated Information Theory (IIT) is a theory of consciousness (Hoel et al., 2013b; Oizumi et al., 2014; Tononi, 2012) that has also been described as an example of 'strong emergence' (Hoel et al., 2013a). The theory's core precept is that a system is conscious if it possesses high levels of a quantity called Φ (phi), which is a measure of the system's capacity of integrating information. Tononi and Sporns (Tononi and Sporns, 2003) argue that this capacity supersedes any other micro-property of the system and is maximally irreducible to its individual components (Hoel et al., 2013a). In other words, IIT equates/conflates consciousness with the emergence of information in the brain surpassing and overriding the information which the brain's constituents already generate independently of one another. If one models the brain as a network of nodes exchanging information with a variety of directed connections, the system will exhibit specialization, if it contains highly connected modules, and integration if modules are highly connected. IIT argues that a high value of Φ for a network can be obtained if the connection patterns of its elements exhibit both high integration and a specialization that leads to activity patterns of the highest complexity from which conscious awareness emerges. Supporting this view, network analyses of fMRI data acquired during deep sleep (N3) indicate increased network modularity compared to conscious wakefulness, suggestive of diminished cortical integration (Spoormaker et al., 2010; Tagliazucchi et al., 2012). Conversely, loss of consciousness also occurs during epileptic seizures when large portions of the cerebral cortex oscillate in synchrony, reflecting abnormally high integration (Blumenfeld, 2012; Cavanna et al., 2017). Although IIT is inspired by cognitive science, it is only weakly constrained by a set of rules/principles that are invariant to the underlying neurobiology. Proponents of IIT have explicitly claimed to go beyond the constituent parts in terms of complexity/information exchange. For ITT, it is not just our current description of the brain that is irreducible to its constituent parts (e.g., because we lack measurement devices, empirical evidence or theoretical tools). Instead, the emergent phenomena are more accurate descriptions of underlying reality (e.g., by providing more accurate cause-effect description) (Marshal et al, 2018).

From Strong to Weak Emergence

The two examples considered above both highlight the two hallmarks of strongly emergent phenomena: (1) emergent phenomena are hypothetically generated from underlying processes and (2) they are somehow autonomous from them. However, this is problematic. Under these conditions, the paradigm of strong emergence seems not to have moved far from the perennial philosophical puzzle of emergent phenomena floating inconsistently over some unspecific physical substrate. The whole of the emergent phenomena still cannot be reduced or explained by

its parts; thus, it follows that no change in its components can have a predictable effect on the whole. If this is the case, it seems reasonable to argue that the science of complex organisms (mereology) is still supported by largely illegitimate metaphysics.

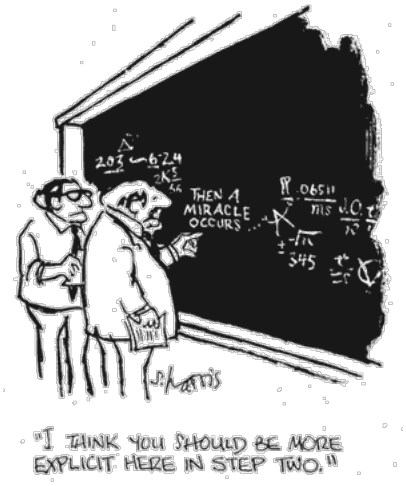

Figure 2: *An illustrative approach to Strong Emergence (Authorized reproduction – S. Harris – Science Cartoon Plus)*

To move the argument further it may be helpful to introduce some more stringent definitions of what actually constitutes strong emergence. The most commonly used are the four principles introduced by O'Connor (O'Connor, 1994):

"A property M is an emergent property of a (mereologically-complex) object O if and only if:

(1) M supervenes on properties of the parts of O; and

(2) M is not had by any of the object's parts; and

(3) M is distinct from any structural property of O, and

(4) M has direct ("downward") determinative influence on the pattern of behaviour involving O's parts."

The problem of the above paradigm is that, in order to fulfil all these requirements, the analysis of emergence becomes rapidly unworkable. In the now classic example by Kim (Kim, 2006), one considers two emergent mental states (M and M∗) that supervene on physical states (Q and Q∗) of object O, respectively. Now it is legitimate to assume that, in the workings of object O, the mental state M is causal to M∗. According to O'Connor's definition, the emergent property M would suffice to explain M∗ and this would not be reducible to any physical state of O. However if M emerges from Q then logically Q is also causal for M∗. If M and Q both explain M∗ then either they are the same thing or the whole paradigm is overdetermined hence implausible (Kim, 2006). Note in fact how, in the example above, the downward determinative influence fades away and physical properties Q and Q∗ become central to the scientific paradigm.

Are we therefore left with the conclusion that emergence is a concept that must be founded within illegitimate metaphysics and unworkable physics? Not necessarily. Good theoretical formulations that explain the underpinning of complexity and emergence have been around for a long time; Herbert Simon already in 1962 was the first to point out that, from an evolutionary perspective, efficient complex systems need be modular, e.g. composed of sub-systems, and they have to be hierarchically organized, e.g. systems are composed of subsystems that, in turn, have their own subsystems, and so on (Simon, 1962). What the complexities of these various natural or man-made systems really are and what output they actually produce is an obviously more complicated question that, at least in part, may be investigated via simulation and generation of complex computational models. For example, recent complex theory has provided an abundance of cellular automata as demonstrable examples of artificial life mimicking the natural order. A cellular automaton is a collection of coloured cells on a grid that evolves through a number of discrete time steps according to a set of rules based on the states of neighbouring cells (Wolfram, 2002). The rules governing the behaviour of the cells are applied iteratively for as many time steps as desired. Cellular automata have notably been used to model the complex dynamics underlying sensory information processing in the human central nervous system (Gobron et al., 2007; Kozma and Puljic, 2013). In this context, the term "emergence" conveys that these automata are able to evolve into complex spatial and/or temporal patterns that may well be unexpected but their formation is straightforwardly deducible from the rules of interaction of the automaton as well as from the initial conditions of the system and its environment. Importantly, the properties of these automata can be determined by observing or simulating the system with a fair amount of calculation but not by any (or at least any simple) process of a priori analysis.

This alternative paradigm, which is significantly more computationally tractable and amenable to analysis, was introduced by Bedau as "weak emergence" (Bedau, 2011; Bedau, 1997). According to his definition:

"*A macrostate M of physical system O with microdynamic D is weakly emergent if and only if M can be derived from D and O's external conditions but only by simulation.* "

In other words, a description is weakly emergent if it can be modelled by a suitable computation and, conversely, computations are metaphysically "weak emergent" only if they contain the simulation of the emergent behaviour from its elementary constituents.

What makes weak emergence especially interesting is its ubiquity. Starting with simple games, such as John Conway's The Game of Life [1](Gardner, 1970) and slime mould dynamics (Reid et al., 2012), the field of complexity science has been studying and developing a great variety of computational as well living and observable models that are by definition "weakly emergent" and

[1] See (https://media.cognitron.co.uk/papers/game-of-life/index.html) for an example of the Game of Life. Note, even from this simple cellular automata model, differences in the initial state and connectivity of the model (See Patterns in the demo for static, dynamic and mechanistically useful examples), result in widely different emergent behavioural dynamics.

are allowing an increased understanding of complex phenomena (Dennett; Miller; West). The output of these models is extraordinary in the sense that they are unexpected, yet allow empirical investigations and comprehensive definitions of their emergent properties, generally in terms of stochastic distributions over a certain defined output space [Figure 2].

Computational Simulation of Brain Function

Theoretically, simulation of weakly emergent systems though the generation of computational models, may be able to encompass some of the phenomena inherent within strong emergent models (Bedau, 2011).

Hence main question is what computational approaches adopted in the modelling of brain function are "weakly" emergent. The literature contains a plethora of mathematical models that have been successful in modelling selective brain functions, from vision (Landy and Movshon, 1991) to working memory (Madl et al., 2015). However, in order to fulfil the weak emergence tenets outlined above, such models should be able to encompass the whole breadth of scales. Strictly speaking, these models should not resort to intermediate pseudo-representations or rely on meta-scale states and dynamics; they should rather be able to link across scales, e.g., from cellular events (i.e. metabolic processes, neurotransmission) to systems-level dynamics to cognition and, ultimately, behaviour. The task is clearly difficult but some whole-brain computational model classes have demonstrated the potential to support this ambitious scientific programme.

One such class of models is coupled oscillators. A remarkable characteristic of this class of models, despite inherent reductive simplicity, is its ability to explain a large variety of phenomena which, regardless of their specific nature and constituents, seem to share common underlying principles that contribute to characteristic biological phenomena such as synchronization (Kuramoto, 1984; Winfree, 1980). In the brain, an oscillator represents the basic cellular computational unit that, at least in the cortex, is composed by the interaction of a pyramidal neuron and a GABA interneuron underpinning basic brain oscillations in gamma frequency (~80Hz) (Borgers and Kopell, 2003, 2005; Tiesinga and Sejnowski, 2009; Whittington et al., 2000) . The tuning of oscillatory activity by glutamate and GABA activity as well as plastic adaptations can be easily parameterized into oscillatory models (Hellyer et al., 2016; Womelsdorf et al., 2014). The effects of other neurotransmitters, such as dopamine and serotonin, can be also incorporated as the differential tuning of the local excitation/inhibition (E/I) ratio (Ciranna, 2006; SŚmiałowski and Bijak, 1987). This basic oscillatory motif seems to replicate at various scales and evidence has been coalescing around the idea that brain activity self-organizes from local neuronal assemblies to cortical structures and lobes (Cabral et al., 2011; Cabral et al., 2014). Models of Kuramoto oscillators with spatial and temporal characteristics of the structural human white matter connectome (Cabral et al., 2014) and analogous variants of Wilson-Cowan mean-field neuronal

models (Deco et al., 2009) or of the Greenberg-Hastings (Haimovici et al., 2013) have been effectively utilized to generate macroscopic brain signals [For an overview of this approach, see Figure 3 A->B]. These are reminiscent of the time-averaged properties of EEG or fMRI data and replicate the dynamical functional connectivity patterns observed empirically (Bhowmik and Shanahan, 2013; Cabral et al., 2014; Deco et al., 2009; Deco et al., 2013; Deco et al., 2017; Ghosh et al., 2008; Hansen et al., 2015). These models could properly incorporate accurate metabolic constrains such as energetic expenditure (Hillary and Grafman; Lord et al., 2013) or plasticity measures (Hellyer et al., 2016). Similarly, the mappings of brain cellular components available either from mRNA (Sunkin et al., 2013) or PET (Rizzo et al., 2016) and noise elements collected from EEG data (Schirner et al., 2018) could be used to improve their biological validity.

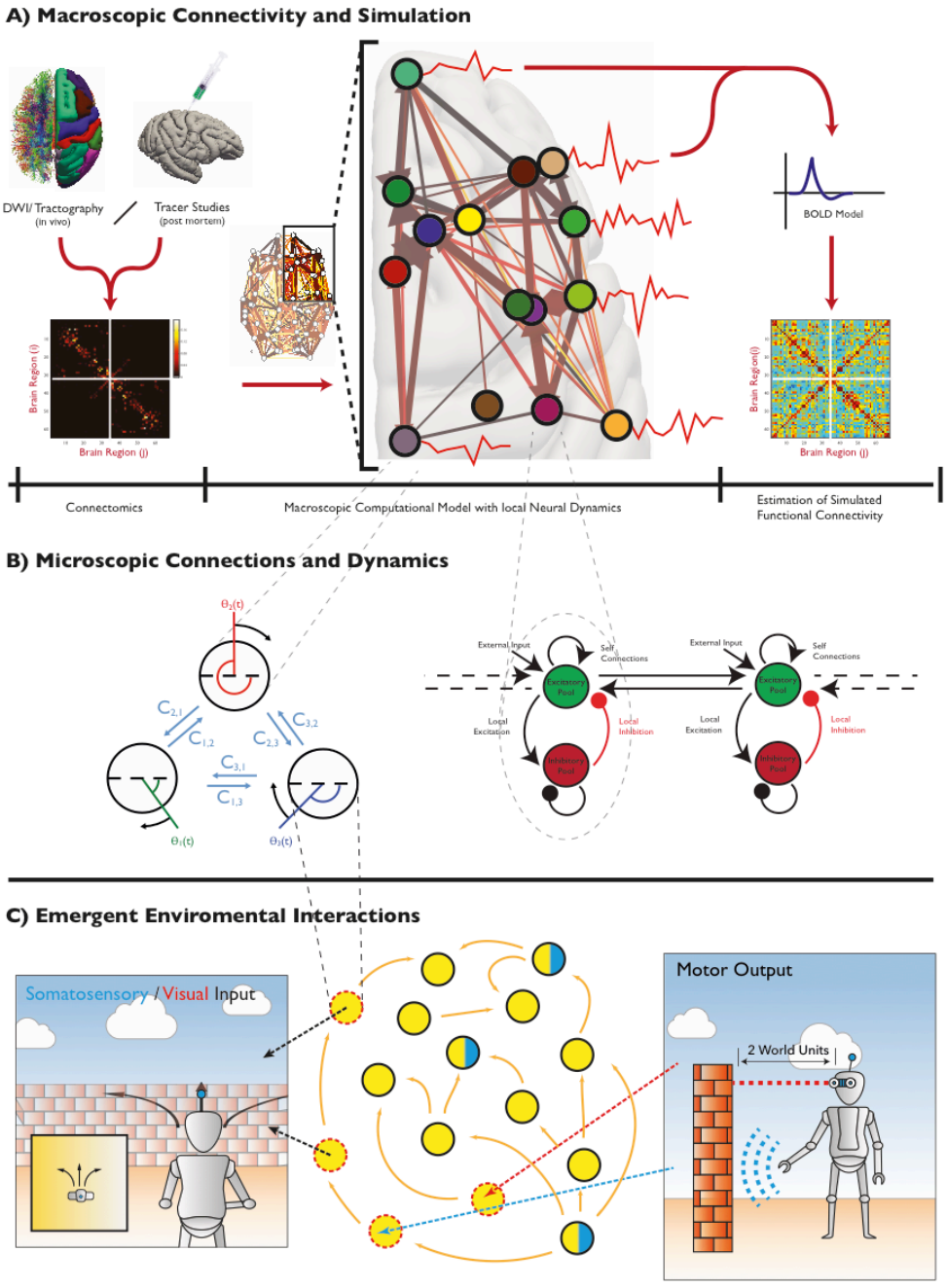

Figure 3: *The use of coupled oscillators to explore emergent properties of neural connectivity at the macroscopic scale.* ***A)*** *The generalised overview of experiments which aims to simulate from structural connectivity of the macroscopic brain, the overall functional activity of a putative neural network – such approaches often generate simulations of fMRI or MEG signals which are then correlated with empirical measures.* ***B)*** *The underlying dynamics at each node, can be simulated using a range of different underlying equations, here we show the simple Kuramoto oscillator system (left), which considers each node as a single reduced phase oscillator, or the more complex (right) Wilson-Cowan model, which exposes for each node, 4 separate interconnections which represent localised (microscopic) connectivity.* ***C)*** *The dynamics of the brain however, do not live in isolation of interactions with the external world, but are a weakly emergent property of this interaction. In our previous work (*(Hellyer et al., 2017)*), we demonstrated one approach for extending exploration of emergent dynamics into the behavioural space – inextricably linking the internal dynamics of the model to their emergent behavioural consequences (Portions of Figure 3, adapted with permission from* (Hellyer et al., 2016)*&* (Hellyer et al., 2017)*)*

Scalability and biological plausibility are important but the key aspect of oscillatory models is that they not only replicate empirical signals but they also seem able to model effectively the emergence of functional properties of the system. For example, network oscillations can tune input selection, temporally aggregate neurons into assemblies, induce synaptic plasticity to create cooperative support of temporal representations and long-term consolidation of information (Beggs, 2008; Buzsaki and Draguhn, 2004; de Arcangelis and Herrmann, 2010; Kinouchi and Copelli, 2006; Moretti and Munoz, 2013; Shanahan, 2010; Shew et al., 2011; Urban et al., 2012). These models have now reached a level of maturity that enables predictions in the clinical realm (Deco and Kringelbach, 2014; Lord et al., 2017; Proix et al., 2017; Zimmermann et al., 2018) from allowing the evaluation of systemic effects of local injuries (Fagerholm et al., 2015) to linking primary sensory and cognitive dysfunctions in schizophrenia (Turkheimer et al., 2015)

The majority of the simulation work so far has investigated spontaneous neural dynamics, in the absence of tasks, sensory input or motor output; the recent literature however has demonstrated the ability of these models to encompass task related activity such as learning and pattern recognition (Capano et al., 2015; van Kessenich et al., 2019). To further inform cognitive sciences, these models have also been constructed to project behaviour. We have recently embodied a computational model of spontaneous neural dynamics into a simulated agent, an avatar, with sensory input from and motor output to a simulated environment (Hellyer et al., 2017) and demonstrated in behavioural terms the results of plastic adaptation of the system. The modelling of interactions between a simulated brain and a simulated environment is still in its infancy but

has demonstrated the potential to explore the emergence of behaviour directly from neurobiologically plausible oscillatory models [Figure 3B->C].

Conclusion

This manuscript is primarily concerned with the epistemological bases of emergence in models of higher cognitive function; this is relevant because the Neurosciences seem to be at an interesting yet familiar junction, reminiscent of the alchemy/chemistry paradigm shift. Despite the remarkable advances brought forward by the labour of healers, artists, clothiers and metal workers for more than 2000 years, it was the introduction of accurate quantitative experiments, explicit analytic thought and experimental verification, combined with an increasing understand of matter that transformed chemistry into a science (Cobb and Goldwhite, 2001).

With the above in mind, we have focused this manuscript on the ontological foundations of wholistic analytical approaches to cognition on the one hand, such as but not limited to FEP and IIT, and, on the other, computational approaches that explore the emergence of higher mental function from the neurobiological micro-scale via simulation, as exemplified by systems of coupled oscillators *in silico*.

We have reviewed the contemporary epistemological literature that suggests that strong emergence e.g. the use of overarching principles to model mental function can be helpful but, without an anchoring to the biological system, provides a merely descriptive tool that for practical (rather than epistemological) reasons is likely to be overdetermined, e.g. too many parameters/explanatory variables will suffice for explanations for the same phenomenon. For example, the Bayesian model of FEP is based on a probabilistic prior that is parameterized by internal states; if these states are not directly and uniquely discernible, so will be the parameters of the priors and the model becomes overdetermined. Overdetermination also undermines IIT when, for example, it postulates the emergence of consciousness out of a particular information flow in the neuronal circuitry inspired by thalamo-cortical circuits and their difference from cerebellar neuronal patterns (Tononi, 2003); as these two anatomical systems differ significantly in a number of other ways, these could equally be postulated as causative to the same phenomenon (Cerullo, 2015). Nevertheless these models would be anchored to a more credible ontology if shown to be valid stochastic approximations of the output of computational renditions anchored to brain biology; for example a Bayesian prior could be modelled out of the computational bottom-up modelling of anticipation (Stephen and Dixon, 2011). In the case of IIT, setting aside the ultimate ontological barrier faced by any model of consciousness (e.g. the “hard” problem of explaining the relationship between physical phenomena, such as brain processes, and personal experience (Chalmers, 1995)), its use of simple models of information flow could be helpful in furthering some intuitive understanding on the computational properties of variously interconnected brain systems (Marshall et al., 2018). Indeed very recent work points to the

amenability of IIT to its embedding in lower level computational models as well as biologically motivated networks (Marinazzo et al., 2014; Mediano et al., 2016; Tagliazucchi, 2017).

On the other hand, the use of neurobiologically plausible, data-testable, generative renderings of higher mental states is ontologically solid but obviously a very challenging proposition. These models require biological fidelity, hence the capacity to be indexed and/or bounded by signalling and metabolic parameters, and they need to be scalable to demonstrate fidelity to brain macro-signals (e.g. EEG, MEG and fMRI) and ultimately to generate credible behaviour.

The incorporation of metabolic constraints, plasticity measures but also the mapping of cellular components available from mRNA studies are just a few of the methods used to introduce biological plausibility into the models (Hillary and Grafman; Lord et al., 2013; Hellyer et al., 2016; Rizzo et at., 2016). In order to do so, the basic element of the computational model needs to be at a micro-level low enough to enable the incorporation of the above-mentioned biological data. For example, in our recent proposal (Turkheimer et al., 2015) the elementary unit was selected as the PING ensemble (pyramidal-interneuron interaction) which is a key determinant of oscillatory activity in the superficial cortical layers, capable of generating beta and gamma oscillations. The modelling of large neural networks using this elementary unit may notably incorporate GABA and glutamate receptor expression data to further tune the local excitability of local neuronal populations into a biologically realistic range.

Importantly, in order to avoid epistemologically "strong" leaps of faith in model construction, future descriptions of cognition should also carefully subdivide behaviour/consciousness in terms of the levels and hierarchies that may be hypothesised to produce a conscious experience (Seth, 2010). Together with the aforementioned, the recent literature provides a number of examples of successful attempts to combine the brain micro and macro signals and brings realistic promise of a viable path for theoretical and computational neuroscience in the coming years.

Finally, what about consciousness itself? Emergence, in its strong version, has been at the core of a number of proposals around consciousness that did not necessarily involve computation per se (for example see (Havlík, 2012) on John Searle theory). Although an informative account of these approaches is outside the scope of this report, it may be useful to report here some words of Bedau on the subject: "*Weak emergence is no universal metaphysical solvent. For example, if (hypothetically, and perhaps per impossible) we were to acquire good evidence that human consciousness is weakly emergent, this would not immediately dissolve all of the philosophical puzzles about consciousness. Still, we would learn the answers to some questions: first, a precise notion of emergence is involved in consciousness; second, this notion of emergence is metaphysically benign. Thus, free from special distractions from emergence, we could focus on the remaining puzzles just about consciousness itself.*" (Bedau, 1997) Hence, while neuroscientists address the problem of consciousness with novel hypotheses and experimental

paradigms (see (Seth, 2010) for a comprehensive account) it may transpire that emergence is not the pivot of a grand theory of consciousness, but just a distraction.

References:

Aristotle, 1994. Metaphysics Books Z and H. Clarendon Press.

Bedau, M., 2011. Weak emergence and computer simulation, in: Humphreys, P., Imbert, C. (Eds.), Models, Simulations, and Representations. Routledge.

Bedau, M.A., 1997. Weak Emergence. Philosophical Perspectives 11, 375-399.

Beggs, J.M., 2008. The criticality hypothesis: how local cortical networks might optimize information processing. Philos.Transact.A Math.Phys Eng Sci. 366, 329-343.

Bhowmik, D., Shanahan, M., 2013. Metastability and Inter-Band Frequency Modulation in Networks of Oscillating Spiking Neuron Populations. PloS one 8.

Blumenfeld, H., 2012. Impaired consciousness in epilepsy. The Lancet. Neurology 11, 814-826.

Borgers, C., Kopell, N., 2003. Synchronization in networks of excitatory and inhibitory neurons with sparse, random connectivity. Neural computation 15, 509-538.

Borgers, C., Kopell, N., 2005. Effects of noisy drive on rhythms in networks of excitatory and inhibitory neurons. Neural computation 17, 557-608.

Breakspear, M., 2017. Dynamic models of large-scale brain activity. Nature neuroscience 20, 340.

Bubic, A., von Cramon, D.Y., Schubotz, R.I., 2010. Prediction, cognition and the brain. Frontiers in human neuroscience 4, 25.

Burns, D.D.R., Sendova-Franks, A.B., Franks, N.R., 2016. The effect of social information on the collective choices of ant colonies. Behavioral Ecology 27, 1033-1040.

Buzsaki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304, 1926-1929.

Cabral, J., Hugues, E., Sporns, O., Deco, G., 2011. Role of local network oscillations in resting-state functional connectivity. NeuroImage 57, 130-139.

Cabral, J., Luckhoo, H., Woolrich, M., Joensson, M., Mohseni, H., Baker, A., Kringelbach, M.L., Deco, G., 2014. Exploring mechanisms of spontaneous functional connectivity in MEG: how delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations. NeuroImage 90, 423-435.

Capano, V., Herrmann, H.J., de Arcangelis, L., 2015. Optimal percentage of inhibitory synapses in multi-task learning. Scientific reports 5, 9895.

Cavanna, F., Vilas, M.G., Palmucci, M., Tagliazucchi, E., 2017. Dynamic functional connectivity and brain metastability during altered states of consciousness. NeuroImage.

Cerullo, M.A., 2015. The Problem with Phi: A Critique of Integrated Information Theory. PLoS computational biology 11, e1004286.

Chalmers, D.J., 1995. Facing up to the problem of consciousness. Journal of Consciousness Studies 2, 200-219.

Ciranna, L., 2006. Serotonin as a Modulator of Glutamate- and GABA-Mediated Neurotransmission: Implications in Physiological Functions and in Pathology. Current Neuropharmacology 4, 101-114.

Clayton, P., Davies, P.C.W., 2006. The re-emergence of emergence : the emergentist hypothesis from science to religion. Oxford University Press, Oxford ; New York.

Cobb, C., Goldwhite, H., 2001. Creations of fire : chemistry's lively history from alchemy to the atomic age. Perseus Pub., Cambridge, Mass.

Corning, P.A., 2002. The re-emergence of "emergence": A venerable concept in search of a theory. Complexity 7, 18-30.

de Arcangelis, L., Herrmann, H.J., 2010. Learning as a phenomenon occurring in a critical state. Proc.Natl.Acad.Sci.U.S.A 107, 3977-3981.

Deco, G., Jirsa, V., McIntosh, A.R., Sporns, O., Kotter, R., 2009. Key role of coupling, delay, and noise in resting brain fluctuations. Proceedings of the National Academy of Sciences of the United States of America 106, 10302-10307.

Deco, G., Jirsa, V.K., McIntosh, A.R., 2011. Emerging concepts for the dynamical organization of resting-state activity in the brain. Nature reviews. Neuroscience 12, 43-56.

Deco, G., Jirsa, V.K., McIntosh, A.R., 2013. Resting brains never rest: computational insights into potential cognitive architectures. Trends in neurosciences 36, 268-274.
Deco, G., Kringelbach, M.L., 2014. Great expectations: using whole-brain computational connectomics for understanding neuropsychiatric disorders. Neuron 84, 892-905.
Deco, G., Kringelbach, M.L., Jirsa, V.K., Ritter, P., 2017. The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core. Scientific reports 7, 3095.
Dennett, D.C., From bacteria to Bach and back : the evolution of minds, First edition. ed.
Fagerholm, E.D., Hellyer, P.J., Scott, G., Leech, R., Sharp, D.J., 2015. Disconnection of network hubs and cognitive impairment after traumatic brain injury. Brain : a journal of neurology 138, 1696-1709.
Friston, K., 2009. The free-energy principle: a rough guide to the brain? Trends in cognitive sciences 13, 293-301.
Friston, K., 2010. The free-energy principle: a unified brain theory? Nature reviews. Neuroscience 11, 127-138.
Friston, K., Kilner, J., Harrison, L., 2006. A free energy principle for the brain. Journal of physiology, Paris 100, 70-87.
Gardner, M., 1970. MATHEMATICAL GAMES. Scientific American 223, 120-123.
Ghosh, A., Rho, Y., McIntosh, A.R., Kotter, R., Jirsa, V.K., 2008. Noise during rest enables the exploration of the brain's dynamic repertoire. PLoS computational biology 4, e1000196.
Gobron, S., Devillard, F., Heit, B., 2007. Retina simulation using cellular automata and GPU programming. Machine Vision and Applications 18, 331-342.
Haimovici, A., Tagliazucchi, E., Balenzuela, P., Chialvo, D.R., 2013. Brain Organization into Resting State Networks Emerges at Criticality on a Model of the Human Connectome. Phys Rev Lett 110.
Hansen, E.C., Battaglia, D., Spiegler, A., Deco, G., Jirsa, V.K., 2015. Functional connectivity dynamics: modeling the switching behavior of the resting state. NeuroImage 105, 525-535.
Havlík, V., 2012. Searle on Emergence. 19, 40-48 %J Organon F: Medzinárodný Časopis Pre Analytickú Filozofiu.
Hellyer, P.J., Clopath, C., Kehagia, A.A., Turkheimer, F.E., Leech, R., 2017. From homeostasis to behavior: Balanced activity in an exploration of embodied dynamic environmental-neural interaction. PLoS computational biology 13, e1005721.
Hellyer, P.J., Jachs, B., Clopath, C., Leech, R., 2016. Local inhibitory plasticity tunes macroscopic brain dynamics and allows the emergence of functional brain networks. NeuroImage 124, 85-95.
Hillary, F.G., Grafman, J.H., Injured Brains and Adaptive Networks: The Benefits and Costs of Hyperconnectivity. Trends in cognitive sciences 21, 385-401.
Hoel, E.P., Albantakis, L., Tononi, G., 2013a. Quantifying causal emergence shows that macro can beat micro. Proceedings of the National Academy of Sciences of the United States of America 110, 19790-19795.
Hoel, E.P., Albantakis, L., Tononi, G., 2013b. Quantifying causal emergence shows that macro can beat micro. Proceedings of the National Academy of Sciences 110, 19790-19795.
Kim, J., 2006. Emergence: Core ideas and issues. Synthese 151, 547-559.
Kinouchi, O., Copelli, M., 2006. Optimal dynamical range of excitable networks at criticality. Nat Phys 2, 348-351.
Kozma, R., Puljic, M., 2013. Hierarchical random cellular neural networks for system-level brain-like signal processing. Neural networks : the official journal of the International Neural Network Society 45, 101-110.
Kuramoto, Y., 1984. Chemical oscillations, waves, and turbulence. Springer-Verlag, Berlin ; New York.
Landy, M.S., Movshon, J.A., 1991. Computational models of visual processing. MIT Press, Cambridge, Mass.
Lestienne, R., 2014. A Bayesian and Emergent View of the Brain. KronoScope 14, 180-193.
Lewes, G.H., 1879. Problems of life and mind. Trübner & co., London,.
Lord, L.D., Expert, P., Huckins, J.F., Turkheimer, F.E., 2013. Cerebral energy metabolism and the brain's functional network architecture: an integrative review. Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism 33, 1347-1354.

Lord, L.D., Stevner, A.B., Deco, G., Kringelbach, M.L., 2017. Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 375.
Madl, T., Chen, K., Montaldi, D., Trappl, R., 2015. Computational cognitive models of spatial memory in navigation space: A review. Neural Networks 65, 18-43.
Marinazzo, D., Pellicoro, M., Wu, G., Angelini, L., Cortes, J.M., Stramaglia, S., 2014. Information transfer and criticality in the Ising model on the human connectome. PloS one 9, e93616.
Marshall, W., Albantakis, L., Tononi, G., 2018. Black-boxing and cause-effect power. PLoS computational biology 14, e1006114.
Mediano, P.A.M., Farah, J.C., Shanahan, M., 2016. Integrated Information and Metastability in Systems of Coupled Oscillators, arXiv e-prints.
Mill, J.S., 1874. A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation, 8th ed. Harper & brothers, New York,.
Miller, J.H., A crude look at the whole : the science of complex systems in business, life, and society.
Mlot, N.J., Tovey, C.A., Hu, D.L., 2011. Fire ants self-assemble into waterproof rafts to survive floods. Proceedings of the National Academy of Sciences of the United States of America 108, 7669-7673.
Moretti, P., Munoz, M.A., 2013. Griffiths phases and the stretching of criticality in brain networks. Nature communications 4, 2521.
O'Connor, T., 1994. Emergent properties. American Philosophical Quarterly 31, 91-104.
Oizumi, M., Albantakis, L., Tononi, G., 2014. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS computational biology 10, e1003588.
Proix, T., Bartolomei, F., Guye, M., Jirsa, V.K., 2017. Individual brain structure and modelling predict seizure propagation. Brain : a journal of neurology 140, 641-654.
Reid, C.R., Latty, T., Dussutour, A., Beekman, M., 2012. Slime mold uses an externalized spatial "memory" to navigate in complex environments. Proceedings of the National Academy of Sciences 109, 17490-17494.
Reid, C.R., Lutz, M.J., Powell, S., Kao, A.B., Couzin, I.D., Garnier, S., 2015. Army ants dynamically adjust living bridges in response to a cost-benefit trade-off. Proceedings of the National Academy of Sciences of the United States of America 112, 15113-15118.
Rizzo, G., Veronese, M., Expert, P., Turkheimer, F.E., Bertoldo, A., 2016. MENGA: A New Comprehensive Tool for the Integration of Neuroimaging Data and the Allen Human Brain Transcriptome Atlas. PloS one 11, e0148744.
Schirner, M., McIntosh, A.R., Jirsa, V., Deco, G., Ritter, P., 2018. Inferring multi-scale neural mechanisms with brain network modelling. eLife 7.
Seth, A.K., 2010. The grand challenge of consciousness. Front Psychol 1, 5.
Shanahan, M., 2010. Metastable chimera states in community-structured oscillator networks. Chaos: An Interdisciplinary Journal of Nonlinear Science 20, 013108.
Shew, W.L., Yang, H., Yu, S., Roy, R., Plenz, D., 2011. Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 55-63.
Simon, H.A., 1962. The Architecture of Complexity. Proceedings of the American Philosophical Society 106.
Sperry, R.W., 1980. Mind-brain interaction: mentalism, yes; dualism, no. Neuroscience 5, 195-206.
Spoormaker, V.I., Schroter, M.S., Gleiser, P.M., Andrade, K.C., Dresler, M., Wehrle, R., Samann, P.G., Czisch, M., 2010. Development of a large-scale functional brain network during human non-rapid eye movement sleep. The Journal of neuroscience : the official journal of the Society for Neuroscience 30, 11379-11387.
SŚmiałowski, A., Bijak, M., 1987. Excitatory and inhibitory action of dopamine on hippocampal neurons in vitro. Involvement of D2 and D1 receptors. Neuroscience 23, 95-101.
Stephen, D.G., Dixon, J.A., 2011. Strong anticipation: Multifractal cascade dynamics modulate scaling in synchronization behaviors. Chaos, Solitons & Fractals 44, 160-168.

Sunkin, S.M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T.L., Thompson, C.L., Hawrylycz, M., Dang, C., 2013. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Research 41, D996-D1008.
Tagliazucchi, E., 2017. The signatures of conscious access and its phenomenology are consistent with large-scale brain communication at criticality. Conscious Cogn 55, 136-147.
Tagliazucchi, E., von Wegner, F., Morzelewski, A., Borisov, S., Jahnke, K., Laufs, H., 2012. Automatic sleep staging using fMRI functional connectivity data. NeuroImage 63, 63-72.
Tiesinga, P., Sejnowski, T.J., 2009. Cortical enlightenment: are attentional gamma oscillations driven by ING or PING? Neuron 63, 727-732.
Tononi, G., 2003. Consciousness: Theoretical Aspects, in: Adelman, G., Smith BH (Ed.), Encyclopedia of Neuroscience. Elsevier, New York.
Tononi, G., 2012. Integrated information theory of consciousness: an updated account. Archives italiennes de biologie 150, 293-329.
Tononi, G., Sporns, O., 2003. Measuring information integration. BMC neuroscience 4, 31.
Turkheimer, F.E., Leech, R., Expert, P., Lord, L.D., Vernon, A.C., 2015. The brain's code and its canonical computational motifs. From sensory cortex to the default mode network: A multi-scale model of brain function in health and disease. Neuroscience and biobehavioral reviews 55, 211-222.
Urban, A., Rancillac, A., Martinez, L., Rossier, J., 2012. Deciphering the Neuronal Circuitry Controlling Local Blood Flow in the Cerebral Cortex with Optogenetics in PV::Cre Transgenic Mice. Frontiers in pharmacology 3, 105.
van Kessenich, L.M., Berger, D., de Arcangelis, L., Herrmann, H.J., 2019. Pattern recognition with neuronal avalanche dynamics. Physical Review E In press.
West, G.B., Scale : the universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies.
Whittington, M.A., Traub, R.D., Kopell, N., Ermentrout, B., Buhl, E.H., 2000. Inhibition-based rhythms: experimental and mathematical observations on network dynamics. International journal of psychophysiology : official journal of the International Organization of Psychophysiology 38, 315-336.
Winfree, A.T., 1980. The geometry of biological time. Springer Verlag, New York.
Wolfram, S., 2002. A new kind of science. Wolfram Media Inc.
Womelsdorf, T., Valiante, T.A., Sahin, N.T., Miller, K.J., Tiesinga, P., 2014. Dynamic circuit motifs underlying rhythmic gain control, gating and integration. Nature neuroscience 17, 1031-1039.
Zimmermann, J., Perry, A., Breakspear, M., Schirner, M., Sachdev, P., Wen, W., Kochan, N., Mapstone, M., Ritter, P., McIntosh, A.R., Solodkin, A., 2018. Differentiation of Alzheimer's disease based on local and global parameters in personalized Virtual Brain models. bioRxiv.