Dynamical Emergence Theory (DET):
A Computational Account of Phenomenal Consciousness

Roy Moyal, Tomer Fekete, and Shimon Edelman

December 2, 2019

Abstract

Scientific theories of consciousness identify its contents with the spatiotemporal structure of neural population activity. We follow up on this approach by stating and motivating Dynamical Emergence Theory (DET), which defines the amount and structure of experience in terms of the intrinsic topology and geometry of a physical system's collective dynamics. Specifically, we posit that distinct perceptual states correspond to coarse-grained macrostates reflecting an optimal partitioning of the system's state space—a notion that aligns with several ideas and results from computational neuroscience and cognitive psychology. We relate DET to existing work, offer predictions for empirical studies, and outline future research directions.

Keywords: perception; vision; awareness; neural dynamics; macrostates; metastability.

> First, conscious awareness, in the present view, is interpreted to be a dynamic emergent property of cerebral excitation.
>
> — Sperry (1969)

1 Terms and Definitions

What does it mean to see? Students of visual perception will remember this question from the opening paragraph of David Marr's book *Vision* (1982), to which he offered the answer "To know what is where by looking." Although the program of Marr and Poggio (1977), which sought to explain perception on the computational, algorithmic, and implementational levels, has led to many successes (Poggio, 2012; Hassabis, Kumaran, Summerfield & Botvinick, 2017), it has also sidestepped the more serious challenge of understanding the relationship between the structure of neural population activity and that of phenomenal consciousness (Edelman, 2016).

In cognitive neuroscience, phenomenal consciousness is typically operationalized in terms of behavioral criteria—for instance, in the case of sensory awareness, the ability to report detection (either verbally or nonverbally; Dehaene, Changeux, Naccache, Sackur & Sergent, 2006). Perceptual phenomena such as binocular rivalry (in which visual awareness is partially suppressed by presenting a different signal to each eye; Tong, Meng & Blake, 2006), in particular, have been utilized in conjunction with neuroimaging and electrophysiology, offering clues as to the qualitative differences between neural population responses that are accompanied by reportable percepts and those that are not (Dehaene, 2014). Alongside these empirical efforts, neurocomputational accounts that aim to derive the necessary and sufficient conditions for phenomenal consciousness from first principles, such as Integrated Information Theory (IIT; Oizumi, Albantakis & Tononi, 2014; Tononi, Boly, Massimini & Koch, 2016) and Geometric Theory (GT; Fekete, 2010; Fekete & Edelman, 2011), have undergone continuous development.

Our account complements and extends these theories to accommodate recent data and promising hypotheses (reviewed in Moyal & Edelman, 2019). Specifically, it brings together several ideas from neuroscience, psychology, and the philosophy of mind, to all of which we are indebted:

- The need for an axiomatic basis or a set of minimal assumptions for a theory of phenomenal consciousness (Tononi, 2008; Fekete & Edelman, 2011; Oizumi et al., 2014).

- The formal isomorphism between the structure of phenomenal consciousness and that of neural population dynamics (Smart, 2004; Spivey, 2006; Edelman, 2008a; Barrett, 2014), both of which are necessarily observer-independent (i.e., *intrinsic* to the system in question; Fekete & Edelman, 2011; Tononi, 2008).

- The emergence of macrostates in the system's activity space (Crutchfield, 1994; Shalizi, 2001; Atmanspacher, 2016; Hoel, Albantakis, Marshall & Tononi, 2016) whose transitions give rise to changes in qualia (which, given the existence of just-noticeable differences in every modality, should be separable; cf. Krueger, 1989).

- The metastability that may characterize these macrostates (Kelso, 1997; Friston, 1997; Freeman & Holmes, 2005; van Leeuwen, 2007; Rabinovich, Huerta & Laurent, 2008; Rabinovich, Huerta, Varona & Afraimovich, 2008; Tognoli & Kelso, 2014; Deco, Kringelbach, Jirsa & Ritter, 2017; Cocchi, Gollo, Zalesky & Breakspear, 2017).

Our focus is thus distinct from that of theories concerned with higher-order awareness (e.g., phenomenological self-models; Metzinger, 2003, 2007, 2018). We seek, instead, a functional and computational understanding of the relationship between the structure of a system's collective dynamics[1] (section 1.2) and that of the basic awareness it is capable of producing (Edelman, 2008a,b; Fekete & Edelman, 2011; Edelman, Moyal & Fekete, 2016)—which, in its minimal form, consists in representations of some aspects of the world (e.g., one's body and its interactions with the environment; Sperry, 1969, 1970). The veracity of any proposed mapping between a system's dynamics and phenomenal content is testable, even if one is only willing to admit a strictly operational definition of awareness.

The core of the thesis we formulate and motivate below is as follows: the contents of phenomenal consciousness are isomorphic to causally effective, coarse-grained macrostates defined over the system's dynamics. These macrostates, more specifically, are the cells of a generating or a Markov partition of the state space (cf. Allefeld, Atmanspacher & Wackermann, 2009; Atmanspacher, 2016). In the wakeful brain, such partitions may be afforded by the itinerant nature of neural population activity, which often exhibits highly coordinated firing patterns punctuated by abrupt, large-scale transitions (a *metastable* regime; Shanahan, 2010; Tognoli & Kelso, 2014). Empirically, one may characterize such transients by examining the geometric and topological structure of a space spanned by measurements of the system's state (e.g., spikes or local field potentials). A review of the neurophysiological data substantiating this approach, with

[1] Because our notion of structure is relational (determined by properties of a system's collective dynamics), it does not rule out physical interpretations or extensions that are non-local.

a focus on the role of long-range thalamocortical coordination in mediating visual awareness and attention, is provided elsewhere (Moyal & Edelman, 2019).

In the remainder of section 1, we define our terms and highlight important theoretical links between the concepts introduced above. In section 2, we spell out our minimal assumptions, state our Dynamical Emergence Theory (DET), and define three measures: representational capacity (which should also reflect the system's level of consciousness; Koch, Massimini, Boly & Tononi, 2016), the *amount* (richness) of experience, and the *nature* (structure) of experience. In section 3, we situate DET in relationship to existing work, provide suggestions for future empirical studies, and outline some of its predictions.

1.1 Multiple Realizability

Our construal of computation is broad (Edelman, 2008b[2]; Fekete & Edelman, 2011). Be it implemented in discrete or continuous systems, computation inheres in the pattern of transitions among well-defined states, whose boundaries may be defined intrinsically or based on externally imposed rules (a crucial distinction that we revisit later). Computation is thus *multiply realizable*, in that a particular operation may be implemented using different physical components. What matters is the functional organization of the system's elements, not their identities, insofar as certain differences in the latter do not alter the former (cf. Chalmers, 1995; Silberstein & McGeever, 1999, p.196ff).

Multiple realizability entails that some properties of the physical substrate of a given computation may not directly reflect its organizational structure. As a trivial example, consider the molecular composition of an organism: though it changes continually during the course of metabolism (e.g., Thompson & Ballou, 1956), the organism's functioning and behavior (i.e., its causal contribution on higher levels of organization) arise from the patterns of molecular interactions. We shall argue here that phenomenal content, similarly, reflects the existence of multiscale structure in the collective dynamics of physical systems whose elements, up to their function, are fungible.

1.2 Structure and Complexity

To dispel any ambiguity, we define a *system* as a set of *elements*, each being a variable (which may correspond to some physical object or volume) whose states can be represented numerically and evolve over continuous time according to a system of differential equations. Though this definition is purposely broad, the mathematical formalism and tools developed in the context of the neural field approach (Coombes, beim Graben, Potthast & Wright, 2014) may be used to further constrain it, with precise implications that cleanly map onto DET (e.g., metastability; Schwappach, Hutt & beim Graben, 2015).

We theoretically define *structure* as the pattern of dependence among the states (actual or measured) of a system's elements in some time interval of interest. Its complexity, which should correspond to the representational capacity afforded by the dynamics[3], may then be defined as a convex (upward) function of the dimension of the state space or the number of degrees of freedom (Atmanspacher, 2016; Moyal & Edelman, 2019). When the system's state space is low-dimensional (e.g., when all neurons fire synchronously or at a consistent phase lag for a prolonged period of time), the richness of its

[2] For a formal definition of computation in systems with continuous dynamics, see Siegelmann and Fishman (1998).

[3] In the case where the representations are intrinsic—that is, arise from the structure of the dynamics (clustering in the state or trajectory space) and are not arbitrarily determined by an external observer (see section 2.1).

representations would be low. That would also be the case when all states are equally likely (dots are evenly spread throughout the space) and the dimension tends to the number of elements. This notion, which is closely related to that of Integrated Information (Tononi, 2008), is central to both DET and IIT.

These provisional definitions suggest a natural (soft or graded) partitioning of any nontrivially structured dynamical system into functional levels of organization, based on the extent to which information about some components' time series is encoded in others. This property is utilized by algorithms that quantify directed causal influence in nonlinear systems, such as convergent cross mapping (Sugihara et al., 2012; Ye, Deyle, Gilarranz & Sugihara, 2015; Clark et al., 2015).

1.3 Emergent Macrostates

How might structure of the kind described in the previous subsection arise in the brain? The concept of *emergence*, which is central to the study of complex systems (e.g., Crutchfield, 1994; Bar-Yam, 2004; Halley & Winkler, 2008; Collier, 2013; Hoel, Albantakis & Tononi, 2013; Ladyman & Wiesner, 2018), is often invoked in biology and cognitive science as a possible answer (Thompson & Varela, 2001; Le Van Quyen, 2003; Rudrauf, Lutz, Cosmelli, Lachaux & Le Van Quyen, 2003; Kauffman & Clayton, 2006; van Leeuwen, 2007; Kirchhoff & Hutto, 2016). In terms of system dynamics, emergence can be seen as the self-organization of functional, effective macrostates over time[4] (Crutchfield 1994). Here, we use the term in the sense of Allefeld and Atmanspacher (2009, p.1) to refer to the mapping that exists between the micro-level description of the system and a particular macro-level, symbolic description that is topologically equivalent to it (Atmanspacher, 2016).

Emergent macrostates can be identified from the dynamics of a physical system through coarse-graining (e.g., Shalizi & Moore, 2003; Hoel et al., 2016), which amounts to the aggregation of a system's states (or state space trajectories) into equivalence classes based on some of their statistical properties. Such a discretization is necessary given the existence of minimally distinguishable qualia (Krueger, 1989). The macrostates, furthermore, must arise out of either a Markov partition (Allefeld et al., 2009) or a generating partition (Kolmogorov, 1958; Sinai, 1959) of the original domain—one in which, all things being equal (that is, without structural changes resulting, e.g., from learning), the boundaries between the macrostates are preserved over time under the system's dynamics (Atmanspacher, 2016). As Shalizi and Moore (2003, p.1) point out, this ensures that the macrostates are self-consistent (stable over time; Harbecke & Atmanspacher, 2012, p.168 and appendix). Thus, the resulting recasting of a system's collective dynamics in terms of emergent coarse-grained macrostates is not merely an observer-relative description thereof (Shalizi and Moore, 2003, p.1).

2 Dynamical Emergence Theory (DET)

2.1 Minimal Assumptions

Having defined these key terms and concepts, we are now ready to spell out the minimal assumptions on which DET is based. Our aim here is to whittle down the list of requirements posed by GT (Fekete & Edelman, 2011; Edelman & Fekete, 2012) and IIT (e.g., Oizumi et al., 2014), retaining only those necessary for explaining how phenomenal representational states map to the collective dynamics of a physical substrate. Specifically, the Inherence and Structure requirements stated below are necessary for enabling a computational account of phenomenality:

[4] A precise definition of self-organization is offered by Shalizi (2004, p.118701-1).

- **Inherence**. The phenomenal experience of a system is an observer-independent property thereof (i.e., is *intrinsic* to it) rather than a matter of outside interpretation or attribution, and so must be any characteristics that define its physical substrate.

- **Structure**. The structure of phenomenal experience—which, fundamentally, reflects discernment among qualia[5]—must be formally isomorphic to that of the system's emergent macrostates and their transitions.

An additional property arises as a consequence of our definition of intrinsic structure (section 1):

- **Effectiveness**. Phenomenal experience is a functional (as opposed to epiphenomenal) trait. In other words, its states and transitions are causally and predictively effective.

We posit that the necessary conditions of Inherence and Structure are also sufficient. In other words, if the pattern of macrostate transitions over the collective dynamics of a physical system implements intrinsic discernment (Fekete & Edelman, 2011, p.807), then the system exhibiting this pattern possesses some degree of phenomenal experience. Importantly, the system's representational capacity and amount (richness) of experience depends on the complexity of this structure (section 2.3).

The Inherence requirement, importantly, rules out digital computation in its familiar form as a candidate medium for phenomenal experience. As pointed out by Fekete and Edelman (2012), the representational states in digital computers are defined by convention—by means of an externally imposed mapping between the values of physical variables and the symbols they stand for—and therefore are not intrinsic to their physical substrate (Tononi, 2008, p.219; Fekete & Edelman, 2011, p.808).

2.2 The Substrates of Experience

In keeping with the paradigm of Marr and Poggio (1977) and with the principle of multiple realizability, we propose to treat the *implementational substrate* (IS) and *computational substrate* (CS) of phenomenal consciousness separately[6].

We define the former (IS) as the collective dynamics of a physical system in a time interval of interest—a segment of its state space trajectory. The latter (CS) is the set of intrinsically structured, causally effective, coarse-grained macrostates of the IS.

Phenomenal experience, then, is the system's trajectory through its set of macrostates (CS). This definition closely follows that of Fekete and Edelman (2011), except that here we argue that the CS-level symbolic dynamics constitutes a functional, emergent property of the IS.

This theoretical move codifies the assumption that distinct physical systems should be capable of experiencing identical qualia if the quasi-discrete macrostates approximated by their dynamics are

[5] Qualia "enable one to discern similarities and differences: they engage discriminations." (Clark, 1985).

[6] We wish to stress that, by separating the implementational and computational substrates (and by appealing to emergence), we do not mean to imply that the latter is somehow nonphysical. Rather, we hold that it is a structural property of the system's collective dynamics or of some underlying physical field (cf. Barrett, 2014).

isomorphic[7] (consider, for instance, two neuronal ensembles—one biological and one artificial—with identical state transition functions). It also alleviates some of the boundary problems that may arise from the commonly perceived need to pick an "objectively" right grain for the dynamics in question (Fekete, van Leeuwen & Edelman, 2016) by appealing to the multiscale nature of both neural population dynamics and phenomenal experience[8].

2.3 Quantifying Experience

Any explicit theory of phenomenal experience should be accompanied by quantitative measures that agree with its structure and capture theoretically interesting aspects thereof (e.g., Balduzzi & Tononi, 2009; Fekete & Edelman, 2011; Oizumi et al., 2014; Fekete et al., 2016; Mediano, Seth & Barrett, 2019). Appropriate measures, in our view, should summarize structural properties of the CS (ideally, ones with clear cognitive counterparts), afford predictive and explanatory power, and be computationally tractable.

Though the space of possible measures is vast, we hold that three basic types are warranted (all defined with respect to a particular time interval): one to quantify the system's overall *representational capacity* (RC), one to quantify the richness or *amount* of experience (AE), and one to characterize the structure or *nature* of experience (NE) and the similarity between (possibly hierarchical) macrostates.

2.3.1 Representational Capacity (RC)

The quantitative definition of representational capacity developed by Fekete (2010), which DET retains, is based on the concept of the trajectory space of a dynamical system—namely, the space of all trajectories (paths through the state space) of a given duration[9] that are possible under the given dynamical regime.

Fekete (2010) posits a ranking of representational states (such as varying degrees of arousal) by their complexity. Specifically, he suggests that linearly combining various measures of complexity derived from labeled data can allow one to define a scalar state indicator function (SIF): higher values of this function are obtained for more complex trajectories, and same-capacity states correspond to the level sets of the SIF[10]. Because representational capacity was defined by Fekete (2010) as the coupling between the complexity of trajectories within a state and the complexity of the structure of the entire (state dependent) trajectory space, the latter was given by the intrinsic structure of these level sets, as quantified by their topological complexity.

Intuitively, the least complex trajectory space is one in which there are no "holes," corresponding to systems in which all trajectories are possible. A system whose dynamics yields a uniformly filled trajectory space (such that any trajectory is allowed) fails to give rise to an intrinsic distinction between different regions of this space (in DET's terminology, macrostates or equivalence classes), and therefore

[7] GT holds that a topological equivalence between the two systems' trajectory spaces should be sufficient (cf. Edelman & Fekete, 2011), but this should be tested empirically to the extent possible.

[8] This corresponds to what William James (1890, p.608) called "the *specious* present": "no knife-edge, but a saddle-back, with a certain breadth of its own on which we sit perched, and from which we look in two directions into time."

[9] This formalization of trajectory spaces is only applicable to trajectories of finite duration.

[10] A level set of a scalar-valued function is a set of points in its domain for which its value is equal to some constant.

between different trajectories (recall section 1.2). The lower bound on the amount of experience is thus zero. In comparison, spaces with nontrivial homological structure are clustered (Fekete, 2010, p.81).

To make the measure of representational complexity sensitive to the expected structure stemming from the hierarchical nature of the perceptual domain's dynamics, a multiscale approach is called for. One such approach is persistent homology, which seeks topological structure that remains unchanged across a certain contiguous range of scales at which it is evaluated (Zomorodian & Carlsson, 2005; Edelsbrunner & Harer, 2008; Fekete et al., 2009).

2.3.2 The Amount of Experience (AE)

We propose to decouple the notion of representational capacity from that of the *amount* (or richness) of experience a system is having in a particular time interval. This move can be justified by observing that the complexity of neural population activity patterns (and the reported phenomenal experiences associated with them) can be changed by adjusting one's sensory input, all without altering one's level of arousal. In vision, for example, the dimensionality of stimulus-evoked EEG patterns is often positively correlated with the geometric complexity of the input (e.g., Müller, Lutzenberger, Preißl, Pulvermüller & Birbaumer, 2003).

Thus, the topological complexity of individual state space trajectories (IS-level) could serve to quantify AE[11]. Though such a measure would inevitably be positively correlated with the system's level of consciousness—ranging from coma to full alertness—and with the topological complexity of its trajectory space (which we used to define RC), it would additionally quantify ongoing changes in the richness of the experienced qualia (e.g., relative to some baseline).

One empirical motivation behind tying AE to the topological complexity of an individual trajectory (of sufficient length) is a theorem due to Takens (1981), which guarantees that the topology of a dynamical system can be approximated from a series of samples of just one of the system's variables. More recently, Deyle and Sugihara (2011) offered a generalized formulation of the theorem, which allows for faster convergence when several variables are tracked. Their convergent cross mapping algorithm (Sugihara et al., 2012), in particular, may allow for a relatively precise characterization of the directed causal influences between components, beyond that afforded by traditional functional and effective connectivity analyses (Friston, 2011). The ability of time-delay embedding to serve as a basis for our AE measure, provided that it scales up from toy examples, is also indicated by the recent work of Garland, Bradley, and Meiss (2016), who used it to reconstruct the topology (specifically, the multiscale homology) of the Lorenz attractor system from time series data.

In this connection, it is interesting to note that it is possible to reconstruct the topology of a dynamical system from "point cloud" (as opposed to temporal sequence) data, if a distance function over points is available (e.g., Singh et al., 2008; Carlsson, 2009; cf. Fekete, 2010 and Fekete & Edelman, 2011). The distance function itself can also be estimated from data (e.g., Talmon & Coifman, 2013; Talmon, Mallat, Zaveri & Coifman, 2015; Talmon & Coifman, 2015; Yair, Talmon, Coifman & Kevrekidis, 2017; Sulam, Romano & Talmon, 2017).

[11] A topologically complex trajectory would correspond to a class of rich ongoing experiences. Note that the precise shape of a trajectory that belongs to such a class is constrained, but not uniquely determined, by its complexity.

2.3.3 The Nature of Experience (NE)

Scalar characterizations of phenomenal experience (RC and AE), on their own, would be incomplete. A third empirical measure is needed that would be structured to a degree and in a manner that match that of experience. Importantly, it should allow one to quantify the *similarity* between the experiences of different systems, or of the same system on different occasions.

There are thus two complementary aspects to the nature of experience (NE). First, given a set of experiences, their similarity function should induce a tree-like structure, akin to that of perceptual and conceptual spaces (which are all "tangled hierarchies"; e.g., Edelman 2008a, Fig. 6.13)[12]. Second, the ongoing structure of experience over time takes the form of a directed graph wherein each node is a macrostate. Since the system may remain in the same macrostate (region of the state space or field configuration) for an arbitrary duration (e.g., when viewing a red dot for 500ms as opposed to 600ms), this information should be indicated by a scalar associated with each node in the graph.

Because there are many ways to define a similarity function that would properly capture these two aspects of NE, we are reluctant to single out a particular one *a priori*. Practically, however, we propose to identify appropriate distance functions over estimated macrostates by testing how well they reflect similarity ratings and just-noticeable differences (cf. Krueger, 1989; Edelman, 1998).

2.3.4 A Note on Measurements

Under the proposed approach, representational capacity and the amount of experience correspond to the topological complexity of the trajectory space or the trajectory itself (respectively) while the nature of experience—with all its idiosyncratic and likely ineffable nuances—corresponds to the structure of the CS-level macrostates and transitions.

We stress that both measures are intended to be empirical; their values must be estimated using a sliding window and are expected to fluctuate as the clique of the system's elements that affect the trajectory of interest changes (e.g., in response to external perturbations). Nevertheless, they should offer insight into the intrinsic topology and geometry of the system, even when the estimation is based on observations that are rather crude in comparison with the actual dynamics (e.g., EEG data). In those cases, we propose to treat the measures as relative, not absolute. Rather than interpret the values they yield for a specific perceptual state, we propose to use them to draw comparisons among several such states, each corresponding to some well-defined and controlled baseline (as suggested in section 3.2).

Now that we have specified the core principles of DET, we proceed to place it in the context of other, similar-scope theories and ideas from neuroscience, list some predictions for empirical studies, and explore avenues for future research.

3 Discussion

DET spans several levels of the Marr-Poggio explanatory hierarchy. On the one hand, our assumptions belong on the computational level. We posit that the problem that phenomenal consciousness is meant to solve is that of intrinsic discernment ("this, not that", as per the Structure constraint), which is inexorably

[12] For example, crimson is a kind of red, which in turn is a kind of color. This structure would be reflected in the clustering of the trajectories.

bundled up with valuation (e.g., the valence of the ongoing experience). This results in intrinsic differential valuation of outcomes and therefore differential predisposition toward courses of action. On the other hand, the key questions we address belong on the algorithmic and implementational levels, insofar as they have to do with system's dynamics (and coarse-grained macrostate transitions) as the physical substrate of experience. It may seem odd to require that a computational theory, which postulates that phenomenal consciousness is multiply realizable, concern itself with the level of physical implementation. The different levels, however, typically constrain each other (Edelman, 2012, p.1122). In this case, since the brain is the only system known to implement awareness, constraints arising from neural organization and function must be imposed.

3.1 DET and Existing Work

The seminal paper by Crick and Koch (1990), which reinvigorated the scientific study of consciousness in the 1990s, did so by encouraging the search for its neural correlates. Though this research program has been successful (Koch et al., 2016), it does not yet explain why particular mechanisms are conscious and others are not. This is also true of mechanistic accounts that identify phenomenality with stable "explicit" representation (O'Brien and Opie, 1999), global dissemination of information (Baars, 2005; Dehaene, King, and Marti, 2014), or convergence to an attractor (Malach, 2012), to single out just a few hypotheses. Why should any of these qualities of brain dynamics necessarily give rise to felt experience? Is every stable or globally shared (Baars, 2005) representation conscious? We hold that, to answer these questions, one must consider the functional role of conscious states, and that is the approach we take by insisting that the dynamics exhibit transitions between causally effective macrostates. Neurocomputational approaches that align with this view are briefly discussed next.

3.1.1 Neurodynamical Frameworks

In response to his early critics, Sperry (1970, p.586) wrote: "The objection that the hypothesis [of dynamical emergence of consciousness] remains vague on details is of course valid and must probably continue to apply […] for some time to come." Decades later, the missing details finally came to be seriously pondered, in papers with titles such as "A cinematographic hypothesis of cortical dynamics in perception" (Freeman, 2006), or "What needs to emerge to make you conscious?" (van Leeuwen, 2007). Several of these neurodynamical perspectives, which attribute a functional role to the transient stabilization of neural activity on multiple spatiotemporal scales (e.g., Kelso, 1997; Friston, 1997; Rabinovich et al., 2001; Kaneko and Tsuda, 2003; Freeman & Holmes, 2005; Tognoli & Kelso, 2013; Rabinovich, Tristan & Varona, 2015), are compatible with DET.

Biological and artificial neural networks, when poised on the "edge of chaos" (Legenstein & Maass, 2007), have been observed to undergo transitions between quasi-stable states (Scarpetta, Apicella, Minati & de Candia, 2018). These states often manifest as alternating periods of increased and decreased synchronization (at different spatiotemporal scales; van Leeuwen, 2007) whose structure consistently reflects particular stimulus properties (Rabinovich, Huerta & Laurent, 2008). Such dynamics are often referred to as a metastable regime (Rabinovich, Huerta, Varona & Afraimovich, 2008; Deco & Kringelbach, 2016)—one in which the system's components constrain each other's states without visiting attractors (Tognoli & Kelso, 2014). Metastable dynamics are characterized by "periods of stable coherence that are themselves inherently unstable" (Friston, 1997), a relative balance between integration and segregation, and the emergence of complex patterns of activity (Tononi, Sporns & Edelman, 1994; Bressler & Kelso, 2001) as the state space trajectory itinerates between structured submanifolds (Friston, 1997). Several teams have provided more specific formulations of this idea. Rabinovich et al. (2001), for instance, define the metastable regime as a series of transitions between saddle fixed points connected by unstable manifolds ("winnerless competition"; Rabinovich, Simmons & Varona, 2015), while Tsuda, Koerner, and Shimizu (1987) propose that neural activity tends to dwell near "attractor ruins" (a behavior dubbed "chaotic itinerancy"; Tsuda, 1991, 1996, 2013, 2015; Kaneko & Tsuda, 2003).

All of these possibilities align with our main thesis—namely, that phenomenal experience can be reduced to symbolic dynamics (CS) by defining intrinsically separable macrostates over the state space of some physical substrate (IS). They also formalize the notion of multiscale spatiotemporal organization (cf. Fekete & Edelman, 2011)—which, as recent electrophysiological data suggest, may manifest as nested oscillations in the thalamocortical network (e.g., Bonnefond, Kastner, & Jensen, 2017). Dynamics of this kind could mediate attentional effects (Edelman & Moyal, 2017): strong perturbations (e.g., high-intensity or behaviorally relevant peripheral input) would dissolve existing phase-aligned population activity patterns and give rise to others. These ideas may be tested more thoroughly by supplementing classic techniques (such as time-frequency analysis and classification) with convergent cross-mapping (e.g., Clark et al., 2015; for determining the functional organization of a system), recurrence analysis (e.g., Marwan, Romano, Thiel & Kurths, 2007; for identifying macrostates), and various time series motif discovery algorithms (e.g., Yeh et al., 2018).

A complete neurodynamical theory of consciousness would not only map the structure of neural activity to a set of well-defined equivalence classes (related by a metric), but also explain *why* such macrostates arise. This latter theoretical goal may be pursued by exploring the relationship between criticality and representational capacity (e.g., Haldeman & Beggs, 2005; Beggs, 2008; Fekete et al., 2018; Moyal & Edelman, 2019). Specifically, it has been proposed that the wakeful brain self-organizes to operate near a phase transition (Beggs & Timme, 2013; Ma, Turrigiano, Wessel & Hengen, 2019), a view supported by studies linking the fine-tuning of certain control parameters (such as neural gain) to the optimization of signal propagation distance and of the ensemble's representational capacity (or the size of its state repertoire; e.g., Haldeman & Beggs, 2005).

3.1.2 Integrated Information Theory (IIT)

Integrated Information Theory (IIT; Tononi, 2008; Oizumi et al., 2014; Tononi et al., 2016) is among the leading computational theories of consciousness. It has originally been formulated for binary discrete dynamical systems; thus, Oizumi et al. (2014, p.4) state that they "[…] consider systems in which the elementary mechanisms are discrete logic gates or linear threshold units […] and assume that these mechanisms are the ones mediating the strongest causal interactions." Attempts to map IIT's formalism to continuous dynamical systems are, for the most part, fairly recent (Hoel et al., 2016; Esteban, Galadí, Langa, Portillo & Soler-Toscano, 2018) and constitute a significant departure from the theory's original formulation. While DET's Inherence and Structure requirements (section 2.1) correspond, conceptually, to IIT's axioms of intrinsic existence and composition, we hold that the other axioms postulated by IIT (integration, information, and exclusion) are in some respects ill-defined, subsumed in the first two, or unnecessary.

The definitions of the quantity and quality of experience offered by IIT (which formalize their notions of integration and information) involve computations over all possible transitions into and out of a momentary state. These are not just intractable (due to the overwhelming combinatorics) but ill-defined in principle, due to the need for a complete knowledge of "all possible" predecessor and successor states (the problem, more specifically, applies to the probability distributions used to define Phi, the amount of consciousness; Barrett and Mediano, 2019, p.4). On the other hand, generating partitions, Kolmogorov-Sinai entropy, and recent generalizations thereof can be defined for ergodic and non-ergodic systems (Barrett & Mediano, 2019), and the empirical measures that DET calls for are tractable.

The conceptual foundations of IIT depend critically on its exclusion principle: "of all overlapping sets of elements, only one set can be conscious—the one whose mechanisms specify a conceptual structure that is *maximally irreducible* (MICS) to independent components. A local maximum of integrated information […] is called a *complex*." (Oizumi et al., 2014, p.9). This applies also to set of mechanisms related by emergence, so that when causal roles of mechanisms on different levels are compared, as they have been by Hoel et al. (2016), the one level whose unique contribution is the largest is considered privileged, to the exclusion of others (this, presumably, is the organizational level on which consciousness resides). Arguably, this notion reflects a conflation of the implementational level ("elements") with the computational level ("conceptual structure"). Since consciousness is a property of the system's collective dynamics (which, physically, could mean the state of some underlying field; Barrett, 2014), attempts to define phenomenal structure (CS) in terms of a particular subset of its physical elements (IS) may be misguided.

Given that discrete dynamics can be approximated by a system whose dynamics on a different (lower) level is in fact continuous, IIT necessarily faces a boundary problem: the need for an intrinsic, objective criterion as to which level of dynamics is the relevant one (Fekete et al., 2016). Arguably, IIT's postulate that only the "complex" exists dismisses all lower-level dynamics. Our preferred alternative to the causal exclusion postulate is the idea of *proportionate causation*, as introduced by Yablo (1992) and discussed by Harbecke and Atmanspacher (2012). We also side with Shalizi (2004), who writes that "coarse-grainings […] are generally multiple levels of more or less detailed descriptions, *all* simultaneously valid for the same physical system."

The last point of comparison between DET and IIT that we touch upon here is their respective definitions of the measure of consciousness (recall section 2.3). According to IIT, "the 'shape' of the constellation of concepts in qualia space completely specifies the quality of a particular experience and distinguishes it from other experiences" (Oizumi et al., 2014, Fig.15). In future work, the IIT notion of qualia space (Balduzzi & Tononi, 2009) can be contrasted with DET's definition of representational structure in terms of the geometry of the system's CS-level trajectory. Furthermore, just like DET distinguishes between the amount and nature of experience, it has been recently proposed that in IIT there should be a distinction between the quantity of experience and its content (Krohn & Ostwald, 2017).

3.1.2 Geometric Theory (GT)

Intrinsic structure of the requisite kind is central to GT (Fekete and Edelman, 2011). While DET can be seen as a direct descendant of that theory, there are also key differences. First, our operational definition of the amount of experience (section 2.3.2) differs from that of representational capacity proposed by Fekete (2010)—crucially, it allows the two to vary somewhat independently, in recognition of the fact that changes in the complexity of neural activity may occur even when the level of consciousness (which imposes an upper bound on the complexity of the intrinsic representations a system may maintain) is kept constant. Second, DET identifies the contents of experience with the dynamics over emergent macrostates (the CS-level)—a question left open in Fekete and Edelman (2011, 2012) and Fekete et al. (2016). Third, DET explicitly allows for the multiple realizability of phenomenal content. Fourth, by acknowledging that causal structure may span multiple levels of organization of the IS, DET avoids sliding down an explanatory slippery slope towards smaller and smaller scales, in search for "the" level where the physics of consciousness is to be found.

3.2 Predictions and Future Directions

In evaluating any computational theory of phenomenality that includes quantitative measures of consciousness, it is important to set realistic criteria for success. First, regarding representational capacity (RC) and the amount of experience (AE), relative rather than absolute values should be of main interest. Second, with regard to the nature of experience (NE), which deals in graph-like structures, a metric needs be defined that would allow for the comparison of such structures in a manner that would match behavioral reports. With these considerations in mind, we proceed to outline some suggestions for empirical studies and future inquiries.

First, we propose to estimate, for a variety of EEG (and perhaps fast fMRI; Grill-Spector & Malach, 2001; Davis & Poldrack, 2013) data, RC as defined by Fekete (2010) and our measures of AE and NE. The results should be compared across changes in the participant's level of arousal (including the rest vs. task distinction) and types of stimulation ("simple" stimuli, such as undifferentiated fields of uniform color, and composite stimuli, such as shapes or scenes of increasing complexity). Though RC and AE should both covary with arousal and reflect the difference between rest and a simple stimulus, the latter should also capture differences between more and less complex sensory stimuli. Differences in NE, as expressed in the representational distance between responses evoked by different stimuli (either in the IS-level state space or per the metric defined on NE graphs) should closely correspond to similarity ratings and other psychophysical measures.

It is also possible to vary the same stimulus along some dimensions (e.g., the orientation or color of a bar). While AE should be invariant to many such changes[13], NE and the empirical metric defined over it should correctly reflect the perceived distance between the presented stimuli as reported by the participant. For complex stimuli that form a controlled pattern in the design space, the configuration formed by NE in the similarity space should reflect the design pattern. This corresponds to the notion of second-order isomorphism (Shepard, 1968; Shepard & Chipman, 1970) between representation spaces and the world (e.g., Cutzu & Edelman, 1996; Edelman, Grill-Spector, Kushnir & Malach, 1998; Edelman, 1998; Op de Beeck, Wagemans & Vogels, 2001).

One may also look for differences in AE and NE between a reference state and a reportable perception state or between states of unawareness and awareness obtained for the same physical stimulus (e.g., using forms of binocular rivalry). Staircase procedures could be used to probe the neural correlates of minimal changes in qualia (in the classic sense of Crick and Koch, 1990). DET predicts that the best correlates (in the explanatory-predictive sense) will be found at the level of emergent macrostates estimated from measurements of local field potentials or spike rates.

Furthermore, we expect that the (high-dimensional) dynamics of the ensemble of neural activities and the (much higher-dimensional) dynamics of the ensemble of synaptic strengths will reflect the same (relatively low-dimensional) space of emergent macrostates and corresponding dynamics, perhaps on

[13] The question of whether AE (the topological complexity of IS-level trajectories) would be invariant to changes in orientation, color, or other visual dimensions is interesting in its own right.

different time scales. As Klopf (1972, p.46) points out, the neuron- and the synapse-based takes on the activity of the brain complement each other. A similar dual view of brain dynamics has been offered more recently by Buzsáki (2010), who coined the concept of "synapsemble" (synaptic ensemble) to complement the concept of an ensemble of neurons and stressed that synaptic weights can also vary on a rapid time scale. In preparations in which synapse-level dynamics can be tracked, it should be possible to apply macrostate identification algorithms (section 2.3) and examine these ideas empirically.

Another prediction that we include here, as an aside, has to do with the evolution of emotions and their relationship to behavior. Consider an evolving population of agents, each endowed with a representation space implemented by an open dynamical system, engaged in sequential behavior, capable of learning in response to behavioral outcomes, and situated in an environment that rewards forethought. We expect the representation spaces for perception and action, harbored by such agents, to evolve trajectory dynamics embodying aversion or tropism with respect to various regions in the representation space.

In this connection, it is interesting to note that the question of what phenomenal consciousness is for (e.g., Campbell, 1974; Thompson & Varela, 2001; Noble, 2008; Cleeremans, 2008; Dehaene et al., 2014; Godfrey-Smith, 2016; Pierson & Trout, 2017) is typically treated somewhat separately from the question of the function and computational nature of emotions (e.g., Sloman et al., 2005; Lowe & Ziemke, 2011; John, Zikopoulos, Bullock & Barbas, 2016; Pessoa, 2017; Bach & Dayan, 2017). Neurodynamical perspectives may be of use in forming a common answer to these two questions. They may also help us understand why (in the functional sense) and how (in the implementational sense) so much of the brain's activity is unconscious.

Finally, one computational mechanism that is capable of implementing metastable dynamics in the service of sequential behavior is competitive queuing (CQ), which has been called upon by cognitive theories of memory, planning, decision making, and language (Houghton & Hartley, 1996; Cooper & Shallice, 2006; Bullock, 2004; Cisek, 2012; Edelman, 2017). The prospect of tying the concept of metastability to CQ is particularly intriguing in light of the postulated causal role of conscious states.

3.3 Conclusion

Dynamical Emergence Theory (DET) aims to map the structure of phenomenal experience to that of the dynamics (actual or observed) of the physical substrates that give rise to it. Minds are, as Minsky (1985) quipped, what brains do; following Sperry (1969) and others, we posit that some such brain doings—specifically, the transitions between coarse-grained macrostates of neural population activity—amount to phenomenal experience. DET modifies the approach of Fekete and Edelman (2011) and complements existing theories, such as Global Workspace Theory (Baars, 2005; Dehaene et al., 2014) and Integrated Information Theory (Tononi, 2008; Oizumi et al., 2014). The theoretical concepts applied here, along with the tentative operational definitions offered, can inform future work on the neural basis of phenomenal consciousness.

Acknowledgments

We thank Reza Shahbazi for comments on an early draft of this paper, Carsten Allefeld for a useful discussion, and the anonymous reviewers for constructive engagement and suggestions.

References

Allefeld, C., Atmanspacher, H., & Wackermann, J. (2009). Mental states as macrostates emerging from brain electrical dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *19*(1), 015102.

Atmanspacher, H. (2016). On macrostates in complex multi-scale systems. *Entropy*, *18*(12), 426.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45-53.

Bach, D. R., & Dayan, P. (2017). Algorithms for survival: a comparative perspective on emotions. *Nature Reviews Neuroscience*, *18*(5), 311.

Balduzzi, D., & Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS computational biology*, *5*(8), e1000462.

Barrett, A. B. (2014). An integration of integrated information theory with fundamental physics. *Frontiers in psychology*, *5*, 63.

Barrett, A. B., & Mediano, P. A. (2019). The Phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2), 11-20.

Bar-Yam, Y. (2004). A mathematical theory of strong emergence using multiscale variety. *Complexity*, *9*(6), 15-24.

Beggs, J. M. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 366*(1864), 329–343.

Beggs, J. M., & Timme, N. (2012). Being critical of criticality in the brain. *Frontiers in physiology*, *3*, 163.

Bonnefond, M., Kastner, S., & Jensen, O. (2017). Communication between brain areas based on nested oscillations. *eneuro*, ENEURO-0153.

Bressler, S. L., & Kelso, J. S. (2001). Cortical coordination dynamics and cognition. *Trends in cognitive sciences*, *5*(1), 26-36.

Bullock, D. (2004). Adaptive neural models of queuing and timing in fluent action. *Trends in cognitive sciences*, *8*(9), 426-433.

Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles, and readers. *Neuron*, *68*(3), 362-385.

Campbell, D. T. (1974). 'Downward Causation' in hierarchically organised biological systems. In F. J. Ayala and T. Dobzhansky (Eds.), *Studies in the Philosophy of Biology*, pp. 179–183. London: Macmillan.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, *46*(2), 255-308.

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

Cisek, P. (2012). Making decisions through a distributed consensus. *Current opinion in neurobiology*, *22*(6), 927-936.

Clark, A. (1985). Qualia and the psychophysiological explanation of color perception. *Synthese*, *65*(3), 377-405.

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., & Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, *96*(5), 1174-1181.

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. In R. Banerjee and B. K. Chakrabarti (Eds.), *Progress in Brain Research, Volume 168*, Chapter 3, pp. 19–33.

Cocchi, L., Gollo, L. L., Zalesky, A., & Breakspear, M. (2017). Criticality in the brain: A synthesis of neurobiology, models and cognition. *Progress in neurobiology*, *158*, 132-152.

Collier, J. (2014). Emergence in dynamical systems. *Analiza i Egzystencja (Analysis and Existence)*, *23*, 17-40.

Coombes, S., beim Graben, P., Potthast, R., & Wright, J. (Eds.). (2014). *Neural fields: theory and applications*. Springer.

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*, 887–916.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. In *Seminars in the Neurosciences*, *2*, 263-275. Saunders Scientific Publications.

Crutchfield, J. P. (1994). The calculi of emergence. *Physica D*, *75*(1-3), 11-54.

Cutzu, F., & Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences*, *93*(21), 12046-12050.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences*, *1296*(1), 108-134.

Deco, G., & Kringelbach, M. L. (2016). Metastability and coherence: extending the communication through coherence hypothesis using a whole-brain computational perspective. *Trends in neurosciences*, *39*(3), 125-135.

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*(1), 3095.

Dehaene, S., Charles, L., King, J. R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current opinion in neurobiology*, *25*, 76-84.

Dehaene, S. (2014). Consciousness and the brain: Deciphering how the brain codes our thoughts. Penguin.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends in cognitive sciences, 10(5), 204-211.

Dehaene, S., Charles, L., King, J. R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current opinion in neurobiology*, *25*, 76-84.

Deyle, E. R., & Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS One*, *6*(3), e18295.

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences 21*, 449–498.

Edelman, S. (2008a). *Computing the mind: How the mind really works*. Oxford University Press.

Edelman, S. (2008b). On the nature of minds, or: Truth and consequences. *Journal of Experimental & Theoretical Artificial Intelligence*, *20*(3), 181-196.

Edelman, S. (2012). Vision, reanimated and reimagined. *Perception*, *41*(9), 1116-1127. Special issue on Marr's *Vision*.

Edelman, S. (2016). The minority report: some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(4), 751-776.

Edelman, S. (2017). Language and other complex behaviors: unifying characteristics, computational models, neural mechanisms. *Language Sciences*, *62*, 91-123.

Edelman, S., & Fekete, T. (2012). Being in time. In S. Edelman, T. Fekete, and N. Zach (Eds.), *Being in Time: Dynamical Models of Phenomenal Experience*, pp. 81–94. John Benjamins.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309-321.

Edelman, S., & Moyal, R. (2017). Fundamental computational constraints on the time course of perception and action. In *Progress in brain research* (Vol. 236, pp. 121-141). Elsevier.

Edelman, S., Moyal, R., & Fekete, T. (2016). To bee or not to bee? *Animal Sentience: An Interdisciplinary Journal on Animal Feeling*, *1*(9), 14.

Edelsbrunner, H., & Harer, J. (2008). Persistent homology-a survey. *Contemporary mathematics*, *453*, 257-282.

Esteban, F. J., Galadi, J. A., Langa, J. A., Portillo, J. R., & Soler-Toscano, F. (2018). Informational structures: A dynamical system approach for integrated information. *PLoS computational biology*, *14*(9), e1006154.

Fekete, T., Pitowsky, I., Grinvald, A., & Omer, D. B. (2009). Arousal increases the representational capacity of cortical tissue. *Journal of computational neuroscience*, *27*(2), 211-227.

Fekete, T. (2010). Representational systems. *Minds and Machines*, *20*(1), 69-101.

Fekete, T., & Edelman, S. (2011). Towards a computational theory of experience. *Consciousness and cognition*, *20*(3), 807-827.

Fekete, T. and Edelman, S. (2012). The (lack of) mental life of some machines. In S. Edelman, T. Fekete, and N. Zach (Eds.), *Being in Time: Dynamical Models of Phenomenal Experience*, pp. 95–120. John Benjamins.

Fekete, T., van Leeuwen, C., & Edelman, S. (2016). System, subsystem, hive: boundary problems in computational theories of consciousness. *Frontiers in psychology*, *7*, 1041.

Fekete, T., Omer, D. B., O'Hashi, K., Grinvald, A., van Leeuwen, C., & Shriki, O. (2018). Critical dynamics, anesthesia and information integration: lessons from multi-scale criticality analysis of voltage imaging data. *Neuroimage*, *183*, 919-933.

Freeman, W. J. (2006). A cinematographic hypothesis of cortical dynamics in perception. *International journal of psychophysiology*, *60*(2), 149-161.

Freeman, W. J., & Holmes, M. D. (2005). Metastability, instability, and state transition in neocortex. *Neural Networks*, *18*(5-6), 497-504.

Friston, K. J. (1997). Transients, metastability, and neuronal dynamics. *Neuroimage*, *5*(2), 164-171.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, *1*(1), 13-36.

Garland, J., Bradley, E., & Meiss, J. D. (2016). Exploring the topology of dynamical reconstructions. *Physica D: Nonlinear Phenomena*, *334*, 49-59.

Godfrey-Smith, P. (2016). Individuality, subjectivity, and minimal cognition. *Biology & Philosophy*, *31*(6), 775-796.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta psychologica*, *107*(1-3), 293-321.

Haldeman, C., & Beggs, J. M. (2005). Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical review letters*, *94*(5), 058101.

Halley, J. D., & Winkler, D. A. (2008). Classification of emergence and its relation to self-organization. *Complexity*, *13*(5), 10-15.

Harbecke, J., & Atmanspacher, H. (2012). Horizontal and vertical determination of mental and neural states. *Journal of Theoretical and Philosophical Psychology*, *32*(3), 161.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245-258.

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1).

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, *110*(49), 19790-19795.

Houghton, G. and Hartley, T. (1996). Parallels models of serial behaviour: Lashley revisited. *Psyche 2*, 25. Symposium on Implicit Learning.

James, W. (1890). The Principles of Psychology. New York: Holt.

John, Y. J., Zikopoulos, B., Bullock, D., & Barbas, H. (2016). The emotional gatekeeper: a computational model of attentional selection and suppression through the pathway from the amygdala to the inhibitory thalamic reticular nucleus. *PLoS computational biology*, *12*(2), e1004722.

Kaneko, K. and Tsuda, I. (2003). Chaotic itinerancy. *Chaos: An Interdisciplinary Journal of Nonlinear*

Science 13, 926–936.

Kauffman, S., & Clayton, P. (2006). On emergence, agency, and organization. *Biology and Philosophy*, *21*(4), 501-521.

Kelso, J. A. S. (1997). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kirchhoff, M. D., & Hutto, D. D. (2016). Never mind the gap: Neurophenomenology, radical enactivism, and the hard problem of consciousness. *Constructivist Foundations 11(2)*, 346–353.

Klopf, A. H. (1972). Brain function and adaptive systems—a heterostatic theory. Technical report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in Proceedings of the International Conference on Systems, Man, and Cybernetics, 1974, IEEE Systems, Man, and Cybernetics Society, Dallas (1972).

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307.

Kolmogorov, A. N. (1958). New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. *Doklady of Russian Academy of Sciences 119*, 861–864.

Krohn, S., & Ostwald, D. (2017). Computing integrated information. *Neuroscience of Consciousness, 3*(1)*,* nix017.

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*(2), 251-267.
Ladyman, J. and Wiesner, K. (2018). *What is a Complex System*. United States: Princeton University Press.

Legenstein, R., & Maass, W. (2007). Edge of chaos and prediction of computational performance for neural circuit models. *Neural networks*, *20*(3), 323-334.

Le Van Quyen, M. (2003). Disentangling the dynamic core: a research program for a neurodynamics at the large-scale. *Biological research, 36*(1), 67-88.

Lowe, R. and Ziemke, T. (2011). The feeling of action tendencies: on the emotional regulation of goal- directed behavior. *Frontiers in Psychology, 2*, 346.

Ma, Z., Turrigiano, G. G., Wessel, R., & Hengen, K. B. (2019). Cortical circuit dynamics are homeostatically tuned to criticality in vivo. *Neuron*.

Malach, R. (2012). Neuronal reflections and subjective awareness. In S. Edelman, T. Fekete, and N. Zach (Eds.), *Being in Time: Dynamical Models of Phenomenal Experience*, pp. 21–36. John Benjamins.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull. 15*, 470–488.

Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, *438*(5-6), 237-329.

Mediano, P., Seth, A., & Barrett, A. (2019). Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy*, *21*(1), 17.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Metzinger, T. (2007). Self models. *Scholarpedia 2*(10), 4174.

Metzinger, T. (2018). Splendor and misery of self-models: Conceptual and empirical issues regarding consciousness and self-consciousness. *ALIUS Bulletin 1*(2), 53–73. Interviewed by J. Limanowski and R. Milliere.

Minsky, M. (1985). *The Society of Mind*. New York: Simon and Schuster.

Moyal, R., & Edelman, S. (2019). Dynamic Computation in Visual Thalamocortical Networks. *Entropy*, *21*(5), 500.

Müller, V., Lutzenberger, W., Preißl, H., Pulvermüller, F., & Birbaumer, N. (2003). Complexity of visual stimuli and non-linear EEG dynamics in humans. *Cognitive Brain Research*, *16*(1), 104-110.

Noble, D. (2008). Claude Bernard, the first systems biologist, and the future of physiology. *Experimental Physiology*, *93*(1), 16-26.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*(1), 127-148.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, *10*(5), e1003588.

Op de Beeck, H. O., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature neuroscience*, *4*(12), 1244.

Pessoa, L. (2017). A network model of the emotional brain. *Trends in cognitive sciences*, *21*(5), 357-371.

Pierson, L. M., & Trout, M. (2017). What is consciousness for? *New Ideas in Psychology*, *47*, 62-71.

Poggio, T. (2012). The levels of understanding framework, revised. *Perception*, *41*(9), 1017-1023.

Rabinovich, M., Huerta, R., & Laurent, G. (2008). Transient dynamics for neural processing. *Science*, 48-50.

Rabinovich, M. I., Huerta, R., Varona, P., & Afraimovich, V. S. (2008). Transient cognitive dynamics, metastability, and decision making. *PLoS computational biology*, *4*(5), e1000072.

Rabinovich, M. I., Simmons, A. N., & Varona, P. (2015). Dynamical bridge between brain and mind. *Trends in cognitive sciences*, *19*(8), 453-461.

Rabinovich, M. I., Tristan, I., & Varona, P. (2015). Hierarchical nonlinear dynamics of human attention. *Neuroscience & Biobehavioral Reviews*, *55*, 18-35.

Rabinovich, M., Volkovskii, A., Lecanda, P., Huerta, R., Abarbanel, H. D. I., & Laurent, G. (2001). Dynamical encoding by networks of competing neuron groups: winnerless competition. *Physical review letters*, *87*(6), 068102.

Rudrauf, D., Lutz, A., Cosmelli, D., Lachaux, J. P., & Le Van Quyen, M. (2003). From autopoiesis to neurophenomenology: Francisco Varela's exploration of the biophysics of being. *Biological research, 36*(1), 27-65.

Scarpetta, S., Apicella, I., Minati, L., & de Candia, A. (2018). Hysteresis, neural avalanches, and critical behavior near a first-order transition of a spiking neural network. *Physical Review E, 97*(6), 062305.

Schwappach, C., Hutt, A., & Beim Graben, P. (2015). Metastable dynamics in heterogeneous neural fields. *Frontiers in systems neuroscience*, *9*, 97.

Shalizi, C. R. (2001). *Causal architecture, complexity and self-organization in the time series and cellular automata* (Doctoral dissertation, University of Wisconsin-Madison).

Shalizi, C. R. (2004). Functionalism, emergence, and collective coordinates: A statistical physics perspective on "What to say to a skeptical metaphysician". *Behavioral and Brain Sciences*, *27*(5), 635-636.

Shalizi, C. R., & Moore, C. (2003). What is a macrostate? Subjective observations and objective dynamics. *arXiv preprint cond-mat/0303625*.

Shanahan, M. (2010). Metastable chimera states in community-structured oscillator networks. Chaos: An *Interdisciplinary Journal of Nonlinear Science, 20*(1), 013108.

Shepard, R. N. (1968). Cognitive psychology: A review of the book by U. Neisser. *Amer. J. Psychol.*, *81*, 285–289.
Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, *1*(1), 1-17.

Siegelmann, H. T., & Fishman, S. (1998). Analog computation with dynamical systems. *Physica D: Nonlinear Phenomena*, *120*(1-2), 214-235.

Silberstein, M., & McGeever, J. (1999). The search for ontological emergence. *The Philosophical Quarterly*, *49*(195), 201-214.

Sinai, Y. G. (1959). On the notion of entropy of a dynamical system. In *Dokl. Akad. Nauk. SSSR* (Vol. 124, p. 768).

Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., & Ringach, D. L. (2008). Topological analysis of population activity in visual cortex. *Journal of vision*, *8*(8), 11-11.

Sloman, A., R. Chrisley, and M. Scheutz (2005). The architectural basis of affective states and processes. In J. Fellous and M. A. Arbib (Eds.), *Who needs emotions? The brain meets the robot*, pp. 203–244. Oxford University Press.

Smart, J. J. (2008). The identity theory of mind. *Stanford Encyclopedia of Philosophy*.

Sperry, R. W. (1969). A modified concept of consciousness. *Psychological review*, *76*(6), 532.

Sperry, R. W. (1970). An objective approach to subjective experience: Further explanation of a hypothesis. *Psychological Review*, *77*, 585–590.

Spivey, M. J. (2006). *The continuity of mind*. New York: Oxford University Press.

Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*, 496-500.

Sulam, J., Romano, Y., & Talmon, R. (2017). Dynamical system classification with diffusion embedding for ECG-based person identification. *Signal Processing*, *130*, 403-411.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980* (pp. 366-381). Springer, Berlin, Heidelberg.

Talmon, R., & Coifman, R. R. (2013). Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proceedings of the National Academy of Sciences*, *110*(31), 12535-12540.

Talmon, R., & Coifman, R. R. (2015). Intrinsic modeling of stochastic dynamical systems using empirical geometry. *Applied and Computational Harmonic Analysis*, *39*(1), 138-160.

Talmon, R., Mallat, S., Zaveri, H., & Coifman, R. R. (2015). Manifold learning for latent variable inference in dynamical systems. *IEEE Transactions on Signal Processing*, *63*(15), 3843-3856.

Thompson, E., & Varela, F. J. (2001). Radical embodiment: neural dynamics and consciousness. *Trends in cognitive sciences*, *5*(10), 418-425.

Thompson, R. C., & Ballou, J. E. (1956). Studies of metabolic turnover with tritium as a tracer. 5. The predominantly non-dynamic state of body constituents in the rat. *Journal of biological chemistry*, *223*, 795-809.

Tognoli, E., & Kelso, J. S. (2013). On the brain's dynamical complexity: coupling and causal influences across spatiotemporal scales. In *Advances in Cognitive Neurodynamics (III)* (pp. 259-265). Springer, Dordrecht.

Tognoli, E., & Kelso, J. S. (2014). The metastable brain. *Neuron*, *81*(1), 35-48.

Tong, F., Meng, M., & Blake, R. (2006). Neural bases of binocular rivalry. Trends in cognitive sciences, 10(11), 502-511.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *The Biological Bulletin*, *215*(3), 216-242.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450.

Tononi, G., Sporns, O., & Edelman, G. M. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, *91*(11), 5033-5037.

Tsuda, I. (1991). Chaotic itinerancy as a dynamical basis of hermeneutics in brain and mind. *World Futures: Journal of General Evolution*, *32*(2-3), 167-184.

Tsuda, I. (1996). A new type of self-organization associated with chaotic dynamics in neural networks. *International Journal of Neural Systems*, *7*(04), 451-459.

Tsuda, I. (2015). Chaotic itinerancy and its roles in cognitive neurodynamics. *Current opinion in neurobiology*, *31*, 67-71.

Tsuda, I., Koerner, E., & Shimizu, H. (1987). Memory dynamics in asynchronous neural networks. *Progress of Theoretical Physics*, *78*(1), 51-71.

van Leeuwen, C. (2007). What needs to emerge to make you conscious? *Journal of Consciousness Studies*, *14*(1-2), 115-136.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245-280.

Yair, O., Talmon, R., Coifman, R. R., & Kevrekidis, I. G. (2017). Reconstruction of normal forms by learning informed observation geometries from data. *Proceedings of the National Academy of Sciences*, *114*(38), E7865-E7874.

Ye, H., Deyle, E. R., Gilarranz, L. J., & Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, *5*, 14750.

Zomorodian, A., & Carlsson, G. (2005). Computing persistent homology. *Discrete & Computational Geometry*, *33*(2), 249-274.