ARTICLE IN PRESS

Consciousness and Cognition xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Towards a computational theory of experience

Tomer Fekete [a,*], Shimon Edelman [b]

[a] Dept. of Biomedical Engineering, Stony Brook University, Bioengineering Building, Room 111, Stony Brook, NY 11794-5281, United States
[b] Dept. of Psychology, Cornell University, 232 Uris Hall, Ithaca, NY 14853-7601, United States

ARTICLE INFO

Article history:
Received 7 July 2010
Available online xxxx

Keywords:
Representation
Experience
Qualia
Computation
State space
Trajectory
Dynamics
Brain activation
Concept
Clustering

ABSTRACT

A standing challenge for the science of mind is to account for the datum that every mind faces in the most immediate – that is, unmediated – fashion: its phenomenal experience. The complementary tasks of explaining what it means for a system to give rise to experience and what constitutes the content of experience (qualia) in computational terms are particularly challenging, given the multiple realizability of computation. In this paper, we identify a set of conditions that a computational theory must satisfy for it to constitute not just a sufficient but a necessary, and therefore naturalistic and intrinsic, explanation of qualia. We show that a common assumption behind many neurocomputational theories of the mind, according to which mind states can be formalized solely in terms of instantaneous vectors of activities of representational units such as neurons, does not meet the requisite conditions, in part because it relies on inactive units to shape presently experienced qualia and implies a homogeneous representation space, which is devoid of intrinsic structure. We then sketch a naturalistic computational theory of qualia, which posits that experience is realized by dynamical activity-space trajectories (rather than points) and that its richness is measured by the representational capacity of the trajectory space in which it unfolds.

1. Introduction

An explanatory framework that treats minds as virtual reality machines – bundles of computations that represent and partly anticipate the statistical structure of the world – is gaining wide acceptance in the cognitive sciences (Churchland & Sejnowski, 1992; Edelman, 2008a; Lindsay & Norman, 1972; Marr, 1982; McDermott, 2001). Within this framework, cognitive computations are construed broadly as dynamical processes that fit certain criteria (Edelman, 2008b). This explanatory move obviates the immaterial distinction between "connectionist" and "classical" computation and consolidates the position of the new computationalism as the overarching theoretical framework for explaining the mind.

By rights, the top priority for the science of mind should be explaining the datum that every mind faces in the most direct and unmediated fashion: its own phenomenal experience, or *qualia*. Following Tye (2007), we construe qualia as "the introspectively accessible, phenomenal aspects of our mental lives." Although intuitively one may distinguish among sensations, concepts, beliefs, thoughts, aspects of being a self, etc., in the end these are all types of qualia. Accordingly, in this paper we gloss over such distinctions and refer interchangeably to the contents of experience, qualia, or simply the phenomenal world.

Whatever else can be said about qualia, one of their most fundamental characteristics is *discernment*: qualia "enable one to discern similarities and differences: they engage discriminations" (Clark, 1985). Had I no capacity for telling apart this ripe tomato's color from other colors (as well as, trivially, from its taste and shape), my experience of red would be greatly diminished; indeed, with all discrimination capacity gone, none would remain of experience either. At the same time,

* Corresponding author.
E-mail addresses: tomer.fekete@gmail.com (T. Fekete), se37@cornell.edu (S. Edelman).

1053-8100/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2011.02.010

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

because a system's experience cannot be a matter of attribution by an external observer (I do not need you or anyone else to mediate my experience of the tomato), any naturalistic theory of experience must explain how discernment, as a necessary attribute of qualia, can be intrinsic to the experiencer. In view of this consideration, we seek to account for the basic fact of experience in terms of a mathematical framework that would show how qualia arise intrinsically from the computational properties of their physical substrate (cf. Tononi, 2008).

Despite the immediacy of phenomenal experience and the urgency of the need to explain its nature, the progress in understanding of qualia, which came to be known as the "hard problem" of consciousness (Chalmers, 1995), has been slow. In other areas of consciousness research, encouraging advances in computational theory have been made in the last few years (Merker, 2007; Metzinger, 2003; see Edelman (2008a), chap. 9, for a synthesis). Although explaining qualia in computational terms has been attempted in the past (O'Brien & Opie, 1999; Smart, 2007; Tononi, 2004, 2008), several key issues remain unresolved.

One of these issues is the lack of an intuitively satisfying and computationally tractable treatment of what we call the *qualia-cline* or *Q-cline* problem (Tononi, 2008). This problem may be stated as follows: Some physical systems (e.g., a rock) clearly cannot have experiences; others (e.g., you, the reader) clearly can; and some (the birds and the bees) seem to fall somewhere in between these extremes. Moreover, it seems prudent to be open to the possibility that as-yet-undiscovered aliens or not-yet-developed advanced robots could also merit a high-Q mark. What is it, then, that differentiates between low-Q and high-Q systems?

Another open issue is explaining the role of *silent units* in shaping experience. Neurobiologists interested in consciousness often equate qualia with momentary states of activation of neural systems or, equivalently, points or vectors in brain state spaces (Edelman, 2002; Smart, 2007). By definition, the location of a point in such a state space is determined by the joint activations of a list of representational units, usually neurons, one per dimension. In real brains, however, the activation is sparse, which means that many (perhaps most) units are silent at any given moment. Equating qualia with momentary states therefore begs the question of how units that happen to be silent at the moment contribute to the present experience or, indeed, to any kind of representation that the system is supposed to harbor.

In the remainder of this paper, we discuss the challenges facing any attempt to explain qualia in computational terms, with a particular focus on the Q-cline and silent units problems. We then sketch an intrinsic computational theory of experience that avoids these problems. The proposed theory rests on two observations. First, construing experience as dynamical and diachronic (extended in time), rather than static and synchronic (momentary), gives voice to silent units and explains how they contribute to the shaping of experience. Second, proper attention to intrinsic representational capacity of activity spaces (i.e., spaces of instantaneous state trajectories) allows one to quantify the sense in which some systems have less rich qualia than others.

2. Dynamics, computation, cognition

Before attempting to explain qualia in computational terms, we briefly recount what we mean by computation. Our account, which follows that of Edelman (2008b), rests on four tenets, as summarized below.

First, every physical process instantiates a *computation* insofar as it progresses from state to state according to dynamics prescribed by the laws of physics, that is, by systems of differential equations. A straightforward way to appreciate this fact is to observe, for instance, that a falling rock computes its trajectory (position as a function of time), simply by obeying Newton's laws of motion.

Second, distinct processes, and even processes whose underlying physics is different, can be governed by the same dynamics. For example, the dynamics of a mass suspended from a spring and of a capacitor-inductor electrical circuit are both captured by the same differential equation. This implies that computation is multiply realizable or, as Chalmers (1994) puts it, it is an organizational invariant.

Third, because of multiple realizability of computation, one computational process or system can *represent* another, in that a correspondence can be drawn between certain organizational aspects of one process and those of the other. In the simplest representational scenario, correspondence holds between successive states of the two processes, as well as between their respective timings. In this case, the state-space trajectory of one system unfolds in lockstep with that of the other system, because the dynamics of the two systems are sufficiently close to one another; for example, formal neurons can be wired up into a network whose dynamics would emulate (Grush, 2004) that of the falling rock mentioned above. More interesting are cases in which the correspondence exists on a more abstract level, for instance between a certain similarity structure over some physical variables "out there" in the world (e.g., between objects that fall like a rock and those that drift down like a leaf) and a conceptual structure over certain instances of neural activity, as well as cases in which the system emulates aspects of its own dynamics. Further still, note that once representational mechanisms have been set in place, they can also be used "offline" (Grush, 2004). In all cases, the combinatorics of the world ensures that the correspondence relationship behind instances of representation is highly non-trivial, that is, unlikely to persist purely as a result of a chance configurational alignment between two randomly picked systems (Chalmers, 1994).

The fourth and final tenet of our computational stance defines a *cognitive* system as a representational system that maintains representations of certain aspects of the world (which can include not only the body but the system itself apart from the "outer" world) and uses these representations to advance its own purposes, such as its continued coherent existence over time, etc. (the notion of "purpose" that applies here is that of Dennett, 2004).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Seeing that computation is central to cognition – both by dint of its role in sustaining representations whereby certain aspects of the world can be captured, processed, and acted upon, and because the problems that cognitive systems must solve are all fundamentally computational (Edelman, 2008a) – the ultimate theory of experience will have to be computational too. What would such a theory look like?

3. Desiderata for a computational theory of experience

Given that cognition is a kind of computation, what special properties, if any, make a cognitive system capable of having phenomenal experiences, and how can one tell that such a system is in the middle of having one, this instant? In this section, we offer a list of constraints that any computational theory of experience should satisfy. These range from metatheoretical, such as explanatory adequacy, to computational, such as categorical and metric properties of the representation spaces realized in the system's activity.

3.1. Fundamental explanatory adequacy

Experience is enabled by physical realization, and therefore its nature is constrained by the nature of the realizing processes. If a process is to realize a given experience, it must have the necessary structure to support its content. Thus, if according to a theory of experience a certain physical state or process realizes phenomenal content, the theory should be explicit as to (i) which properties and features of such physical processes express experiential content, and (ii) complementarily, how systematic change in these properties and features affects the realized experience. For example, if it is the firing of neurons that realizes experience, how is it that some kinds of neural activity do not give rise to experience (e.g., seizure), or that active brains vary greatly in the richness of the experience they realize at different times (dreamless sleep vs. alert and attentive wakefulness)? Likewise, if experience is posited to result from gamma-band oscillations in a population of neurons, the theory should explain how those oscillations express content and how content decays or disappears as the oscillatory activity changes over time.

This line of reasoning suggests that a theory of qualia would have to contain a notion of *representational state,* corresponding to state of vigilance, or, more generally, state of consciousness. This notion would enable one to quantify how changes in structural properties of activity in the representational substrate (neural in the case of the brain) lead to change in the richness or quality of experience. We note that such a theory would necessarily include criteria for demarcating information that is represented unconsciously from information that constitutes the content of experience.

One of the pitfalls in equating the structure of a physical representational process with the structure of experience is conflating between structural facets of the representational medium as such (e.g., certain aspects of the brain's structure or activity), and its structural facets *qua* realizing experience. Consider, for example, the map of the human body found in the primary somatosensory cortex – the so-called somatosensory homunculus discovered by Penfield. For the most part, this map follows the spatial organization of the body: the representation of the knee is next to that of the shin area, and so on. It is tempting to conclude that this organization is in the root of our experience of the spatial organization of our bodies. However, there are a few notable exceptions to the orderly spatial layout of the somatosensory homunculus. For example, the area of the map corresponding to the toes is adjacent to that corresponding to the genitalia, a neuroanatomical quirk that has no counterpart in actual body structure or, for that matter, in immediate experience.

The lesson here is that while some kind of isomorphism between brain processes and the content of experience is a necessary condition for realizing experience, it is by no means a sufficient condition. Only within the broader context of a developed computational theory could an apparent isomorphism (or correlation) be elevated to the vaunted explanatory status we are seeking after. Thus, until shown otherwise, it must be assumed that orderly brain maps result from various mundane constraints such as wiring optimization or sensorimotor coordination, rather than that they are telltales of constraints on phenomenal experience.

3.2. Reasonable scope

One of the most interesting benefits of a viable theory of experience is that it demarcates the boundaries of consciousness. Theories of consciousness range from neural chauvinism (only neurons can realize experience) to panpsychism (any kind of organized matter has qualia to some extent). It seems to us that a reasonable explanatory framework should be highly exclusive, while being general enough to include non-human (and possibly machine) intelligence. A computational framework is by definition multiply realizable (Bickle, 2006; Chalmers, 1994), hence is more in danger of being too inclusive. Therefore, to be plausible it must be accompanied by strict conditions on implementation (to forestall, among other things, the objection from funny instantiation (Maudlin, 1989), used for example by Searle in his "Wordstar" argument (1990)).

The multiple realizability of computation can be formalized through the notion of *functional equivalence*. Functionally equivalent systems are systems that instantiate the same computational structure. A computational theory of experience must explain how two functionally equivalent systems (agents) can share similar experience. The point here is not to engage in hair-splitting philosophical arguments, such as whether or not a Martian's experience of his manganese smoothie is the same as my experience of *this* slice of pizza. Rather, functional equivalence should hold first and foremost between the

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

system and itself (at different times, in different contexts, etc.). In the strict sense, I am not identical to my earlier self (Hume, 2007(1748), Mach, 1886). In particular, each time I learn something new, I undergo structural changes (which in terms of a computational scheme are changes to the parameters of the equations governing the system's dynamics). Are all my experiences thereby irrevocably transformed (cf. Dennett, 1988)? Is it reasonable to maintain that my identity has nevertheless been preserved? What happens in case of massive brain damage? Are qualia simply diminished in proportion to the amount of damage, or are they no longer commensurable? A viable theory of experience should include explicit (and convincing) measures to quantify these issues, which by extension would also serve to account for the possibly similar experiences shared by different systems: other people, different species, and so on.

3.3. A range of levels of granularity

A successful theory of qualia must explicitly account for the conceptual structure of experience. While concepts surely do not capture all the detail and richness of experience,[1] they nevertheless reflect the higher-order organization of the underlying phenomenology. How can a brain-based theory of qualia do so? On the one hand, the workings of the brain at the neuronal level are paramount to the content and structure of experience. On the other hand, the staggering discrepancy between the nominal dimensionality of neuronal activity patterns and the phenomenal dimensionality of conceptual domains suggests that conceptual structure is to be found in some higher-order organization of the underlying activity at the cellular (and sub-cellular) level.

This implies that a comprehensive computational theory of experience should stretch from the level of neuronal activity right up to the conceptual structure of experience, by explicating meaningful and measurable quantities and invariants pertaining to each level of the computational hierarchy realized by brains. Referring to such a hierarchy as "emergent" amounts to shrugging off the explanatory burden, unless it is stated how and when higher-order features of experience are instantiated. If a certain high level property can emerge from a low level one, what are the conditions for that to happen? How do different high level properties and quantities interact and relate to each other? Are they mutually exclusive within a model? Do they require a narrow range of model parameters (which would amount to a critical prediction)? Can they be estimated without measuring the activity of the system all the way to the bottom?

Conceptual structure that emerges from experience facilitates learning, because it makes experience amenable to a coarse-grained and therefore computationally tractable re-representation and reuse (Edelman, 2008b). Such structure, however, is a property of an entire system, which is not present as such in any of the system's components. Accordingly, construing the representational tasks carried out by single neurons (or parts of the brain) in terms of experience realized by the system in its entirety amounts to a category mistake.

It is worth revisiting in the light of the above observation the familiar trope of grandmother neuron (Barlow, 1972). A true grandmother cell would fire when encountering the actual object of its affection, regardless of what Grandmother is wearing, the style of her hair or whether or not she is smiling. At the same time, unless we assume that other cells, which are tuned, e.g., to nose, legs, eyes, and so on, are also firing (in an appropriate pattern), our hypothetical grandmother cell would be very much out of touch with one's experience of the grandmother. In this example, successful representation of the concept "grandmother" requires the coherent representational efforts of numerous representational units, signaling various features and properties necessary to support the experience of a grandmother, which a unit tuned exclusively to the abstraction such as "grandmother" cannot support with its activity alone. Note that a similar argument also holds for any assumed feature detector, regardless of its simplicity, because to form a coherent picture, wholes should match parts, and parts should match not only wholes but other parts as well.

3.4. Counterfactually stable account of implementation

To claim a computational understanding of a system, it is necessary for us to be able to map its instantaneous states and variables to those of a model. Such a mapping is, however, far from sufficient to establish that the system is actually implementing the model: without additional constraints, a large enough conglomerate of objects and events can be mapped so as to realize any arbitrary computation (Chalmers, 1994; Putnam, 1988). A careful analysis of what it means for a physical system to implement an abstract computation (Chalmers, 1994; Maudlin, 1989) suggests that, in addition to specifying a mapping between the respective instantaneous states of the system and the computational model, one needs to spell out the rules that govern the causal transitions between corresponding instantaneous states in a counterfactually resistant manner.

In the case of modeling phenomenal experience, the stakes are actually much higher: one expects a model of qualia to be not merely good (in the sense of the goodness of fit between the model and its object), but true and unique.[2] Given that a

[1] Nor could they: given that conceptual domains realized by actual systems must be finite, this would imply that at the base of this structure there are categories that are conceptually opaque. Also, ultimately, as the ancients phrased it, a concept is one over the many. Thus, different instantiations of a concept, i.e., particular experiences, by definition have qualities over and above their conceptual roles.

[2] As an example of how the requirement of uniqueness can be satisfied by a mathematical model, consider Edelman's (1999, Appendix C) definition of deformable shape representation space based on the theory of Riemann surfaces. Briefly, in this model two shapes of the same topological genus are considered equivalent if related by a conformal mapping; a quasiconformal mapping takes one shape class to another. The resulting shape space, known as the Teichmüller space, has a natural Riemannian metric, in which distance (dissimilarity) between two shapes is defined by deviation from conformality of the quasiconformal mapping by which they are related. Crucially, in Teichmüller theory the parameterization of Riemann surfaces in a given equivalence class of (conformal) deformations is *unique*, as implied by the solution to the Problem of Moduli in algebraic geometry (see Edelman, 1999, pp. 273–274 for references).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

multitude of distinct but equally good computational models may exist, why is not the system realizing a multitude of different experiences at a given time? Dodging this question amounts to conceding that computation is not nomologically related to qualia.

Construing computation in terms of causal interactions between instantaneous states and variables of a system has ramifications that may seem problematic for modeling experience. If computations and their implementations are individuated in terms of causal networks, then any given, specific experience or quale is individuated (in part) by the system's entire space of possible instantaneous states and their causal interrelationships. In other words, the experience that is unfolding *now* is defined in part by the entire spectrum of *possible* experiences available to the system.

In subsequent sections, we will show that this explanatory problem is not in fact insurmountable, by outlining a solution for it. Meanwhile, we stress that while computation can be explicated by numbering the instantaneous states of a system and listing rules of transition between these states, it can also be formulated equivalently in dynamical terms, by defining (local) variables and the dynamics that govern their changes over time. For example, in neural-like models computation can be explicated in terms of the instantaneous state of "representational units" and the differential equations that together with present input lead to the unfolding of each unit's activity over time. Under this description, computational structure results entirely from local physical interactions.

The blurring of the distinction between the possible and the actual has driven some scholars to question the suitability of the computational approach to modeling experience. This concern is, however, unfounded: the now classical analysis by Quine (1951), which revealed linguistic meaning to be holistic,[3] extends naturally to the content of experience. Thus, it is only reasonable that holism should also turn out to be an essential property of any purported computational framework for modeling experience. In fact, anything less than holism on the part of a theory would imply that that it is failing in its job of capturing the essence of experience.

3.5. The structure of activity space and representational capacity

We refer to the space of possible states of a system as its *activity space*. As noted earlier, while computation has what it needs to model experience, it must be constrained to do so properly. The molecules that comprise a mere rock instantiate jointly a very complex activity space simply by following the laws of physics, as everything in the universe does. The burden upon computational theories of experience is to explicate both (i) what features of activity spaces make them appropriate for realizing experience and (ii) how the composition of a given system structures its activity space.

One of the crucial aspects in assigning activity a formal role in explaining phenomenal experience is whether activity is construed in the classical sense of dynamical systems (i.e., as the instantaneous state of the system), or as a process extended in time, that is, as a *trajectory* in the system's instantaneous state space (Churchland, Cunningham, et al., 2010; Churchland, Yu, et al., 2010; Churchland et al., 2007; Geffen et al., 2009; Mazor & Laurent, 2005; Yu et al., 2009). To observe this distinction, we will refer either to instantaneous activity space or to activity trajectory space where the distinction is called for, and will otherwise use the non-committal expression "activity space." The functional difference between instantaneous states and trajectories will be the subject matter of Section 4 and parts of Section 5.

As we argued above, a fundamental constraint on the organization of the activity space of an experiential system is suitability for capturing conceptual structure: insofar as qualia reflect concepts, the underlying activity must do so as well. The basic means of realizing conceptual structure is *clustering* of activity: a representational system embodies concepts by parceling the world (or rather experience) into categories through the discernments or distinctions that it induces over it.[4] As it gives rise to experience, qua instantiating qualia, activity should possess no more and no less detail than that found in the corresponding experience. Specifically, activities realizing different instances of the same concept class must share a family resemblance (Wittgenstein, 1953), while being distinct from activities realizing different concepts. This means that the activity space must divide itself intrinsically into compartments, structured by the requisite within- and between-concept similarity relations.

The crucial point here is that within this framework the richness of the experience realized by a system corresponds to the degree to which its activity separates itself into clusters. The reason is simple: the more clustered the system's activity, the more distinctions it can draw. Moreover, activity being *the* realization of experience, it is not supposed to require any further interpretation. In other words, activity must impose structure on experience intrinsically, or not at all. Accordingly, if a system does not exhibit intrinsically clustered activity, it cannot be engaging in the representation of its environment in any interesting way, as its activity does not in itself induce any distinctions, and hence its phenomenal field (i.e., everything that makes up its awareness at a given moment) remains undifferentiated. Consider a system that gives rise to a homogeneous activity space: say, its activity is equally likely to occupy any point inside an n-dimensional cube (n being the number

[3] In support of holism, one may also quote James (1976(1912), p. 25, "The peculiarity of our experiences, that they not only are, but are known, which their 'conscious' quality is invoked to explain, is better explained by their relations – these relations themselves being experiences – to one another") and Lashley (1923, p.261: "Indeed one must say that a single element never is known except in combination with others. The essence of consciousness is a field of many elements, organized after the plan of human experience.").

[4] In terms of experience, distinctions made at the operational level are manifested as differentiation in the phenomenal field (everything that makes up awareness at a given moment). If, say, two different odorants evoke indistinguishable qualia, the underlying activities must have been indistinguishable (in the metric sense) as well.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

of degrees of representational freedom of the system). Such a homogeneous volume in itself does not suggest any partitioning, and any division of it into compartments would be arbitrary. Thus, the activity of this system cannot amount to experience.

Various subtler distinctions concerning the structure of clusters can be made (and quantified, as we will describe below). One important issue here is the hierarchical structure of clusters (clusters of clusters and so on). In the case of conceptual structure, hierarchy is a means of realizing dominance or inclusion relations among concepts. Other important relations can be modeled by the spatial layout of clusters in the activity space. For example, visual objects can be distinguished according to several parameters such as shape, color, texture, etc., which may be represented by various dimensions of the activity space. Similarly, subdomains of conceptual structures may vary in their dimensionality. Accordingly, the local *effective dimensionality* of configurations of clusters in the activity space is crucial in realizing a conceptual domain.

3.6. Metric properties of models of experience

Articulating a comprehensive computational theory of experience implies that a concrete instantiation of it (in terms of the parameters of a particular model) results in the activity space of the system realizing a particular conceptual domain.[5] We do not contend that models absolutely must be specified to an exhaustive level of detail – i.e., specifying the semantics of the system to the level of correspondence between specific activity and qualia – to be viable, or useful for medical or engineering applications. Still, a good model of experience should specify experience at *some* level of granularity (e.g., loss of consciousness and graded richness of experience must be accounted for). The plausibility of any such model depends in no small measure on how tweaking its structure and parameters affects the realized semantics and conceptual domain. Thus, we require that models be semantically robust in a metric sense. This means that modeling takes place in a space of possible systems, and that models behave smoothly within that space and attain desired properties in clear-cut scenarios.

For example, if systematic changes in structure and parameters of a model lead the resulting system into a null semantics (no experience), they should do so gradually; e.g., in a brain-like model, a gradual decrease in the number of representational units should lead to a gradual decay in the richness of the structure of activity space (and the richness of the corresponding experiences). If not, slight structural damage (corresponding, for example, to the commonplace occurrence of neurons dying) would result in drastic changes in the realized system. In such a model, functional continuity of individuals over time would be unattainable. Similarly, in modeling important cognitive phenomena such as learning, gradual changes in structural aspects of the model should be "semantically plausible." Of course, a model may induce an occasional catastrophic state transition in the corresponding system space. Rather than ruling out a model off the bat, such occasions should be seen as critical predictions for an experimental investigation.

In addition to continuity, models can be characterized by their locality: the degree to which changing local aspects of the model (e.g., its response to a specific initial condition or input) has an effect on the realized activity space at large. No matter how it is quantified, locality is likely to have a crucial impact on the holistic properties of model semantics.

3.7. Computational viability

A computational model is viable not only to the extent that it explains and predicts actual data, but also to the degree to which it enables constructing artificial systems worthy of comparison with the human mind. Therefore such a model should be not only concisely statable, but computationally tractable (in the usual sense defined in computational complexity theory).

4. The shortcomings of instantaneous state-vector semantics

In theories of experience that adopt the instantaneous state-vector semantics, each transient state of the system is uniquely associated with a specific quale, or a fleeting experience. The appeal of state-vector semantics seems to be universal: it is ubiquitous in computational theorizing and philosophical analysis of experience (Brown, 2006; Smart, 2007) and is tacitly assumed in neuroscientific research into consciousness (Crick & Koch, 1990). The attractiveness of this approach seems to stem from a deep-rooted constructivist intuition, according to which experience can be woven frame by frame, just as phenomenally continuous movies are constituted by a succession of brief images. As we shall see next, careful scrutiny in light of the observations that we made so far shows that this theoretical approach to experience is flawed beyond repair.

4.1. The problem of causal structure

Because it is what ultimately realizes experience, the activity of a representational medium should suffice to establish mental content. Another way of putting it is that activity needs no further interpretation to result in experience: anything else leads to infinite regress. However, as noted above, a model that realizes a computational structure must specify not only

[5] This is not to say that on doing so the modeler would become privy to the semantics of the modeled system. Rather, one would have to rely on analogy to known systems (which must be sufficiently similar to the target system), or on extrapolation from the target's self-reports.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

viable transient states, but also the causal mechanism determining the transitions between them. Specifying an instantaneous-state-to-quale mapping alone leaves us with two options: either it is not by virtue of computation that transient states acquire their content (meaning), or casual structure is somehow tacitly implied by the structure of instantaneous patterns of activity.

Let us consider the first possibility. If the meaning of an instantaneous state is independent of computation (according to the notion of computation adopted in Sections 2 and 3, which stated what it means for a system to implement a computation), then transient states have meaning regardless of the mechanism instantiating them. This of course would result in the most extreme kind of panpsychism imaginable, namely, flickers of conscious experience arising spontaneously in any constellation of matter whatsoever, including astronomical numbers of sub-patterns existing within any given pattern of such experience, all at the same time.

We are thus left with the second one, namely, that transient activity patterns somehow carry within them the causal constraints that force their orderly instantiation. Alas, this idea is quite foreign to dynamical systems theory: while the causal mechanism, i.e., the dynamics of the system, surely constrains the orderly progression of transient states, the states in themselves do not constrain the causal structure of the dynamics.

4.2. The problem of silent units

Suppose that we identify the phenomenal experiences of a cognitive system with points in a representation space spanned by a set of units, each of which is either active or not at any given instant of time (as in a network of spiking neurons). Under such a model, the experiences of the system would be constituted by an ever-changing set of units (depending on which units are active at the given time), raising the question of what is it that each inactive (silent) unit contributes to experience, if anything. If the activation in the system is very sparse (as it is the case with higher-level visual areas in the primate brain, for example), the conundrum becomes extreme, but it is troubling enough even if only a few units remain at times silent.

One may try to shrug this problem off by resorting to an atomist approach, under which each representational unit – in the case of brains, each neuron – makes a small, independent, and additive contribution to experience, akin to that of a pixel in an image. Alas, this approach fails in the case of brains: both in slow wave sleep (which is mostly dreamless) and in various types of seizures (especially absence seizures), neural activity is strong and widespread, yet it is not accompanied by massive experience. This suggests that it is the *pattern* of firing that matters here.

The problem of silent units is very general: it holds for any approach to cognition that attributes any kind of content at all to the instantaneous state of the system in question. This includes, for example, the relatively innocuous and uncontroversial notion that an instantaneous state of a system *represents* some state of affairs outside it, an idea that is routinely taken for granted by cognitive scientists (including an early version of one of the present authors; Edelman, 1999, 2002).

Returning to the issue of explaining experience, the problem of silent units holds, for example, for the otherwise intriguing and elaborate Information Integration scheme of Tononi (2008). He seems to be aware of the problem, which, however, he presents not so much as an explanatory lacuna as a fascinating empirical prediction: if silent units indeed contribute to a subject's phenomenal experience, cooling some of them reversibly (which would inactivate them without causing them to fire) would alter the experience. Having offered this prediction, Tononi (2008) stops short of explicitly addressing the crux of the problem: how can presently inactive – silent or silenced – units contribute to *present* experience? As he puts it, "consciousness can be characterized extrinsically as a disposition or *potentiality* – in this case as the potential discriminations that a complex can do on its possible states, through all combinations of its mechanisms." It seems to us that making experience depend on a *potentiality* without explaining how it *actually* comes to pass falls short of completing the explanatory move.

4.3. The problem of activity-space footprint

By restricting the scrutiny of a representational system exclusively to its instantaneous configurations, one may get a misleading notion of what the system actually represents and what, if anything, it is experiencing. The reason is that instantaneous snapshots of the dynamics of a system leave its interpretation too unconstrained, the only constraint on the possible patterns being that individual variables attain values within their respective bounds, due to the physical limitations of the components of the system. The locus of instantaneous states conforming only to this constraint is an n-dimensional cubical volume, implying that in principle the system can attain every possible transient state within it. Thus, models based solely on instantaneous, transient states are characterized by a homogeneous activity space. Because there is no natural *intrinsic* way to partition a homogeneous space, such models fail to realize a cluster (concept) structure. This makes them ill-suited to support experience.

It could be argued that instantaneous states in such a model may not be equiprobable, implying that if the system keeps track of its own deviations from equiprobability, structure could emerge from the mess. For that, however, the system would have to further interpret its own transient states in order to sanction them as expressions of experience. This, of course, would lead to infinite regress: experience *inheres* in activity and therefore any interpretative process leading to it would be manifested in the structure of activity space, which therefore would have to possess intrinsic structure to begin with.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

4.4. The problem of the signature of experience

In brain activity (as measured, e.g., electrically), the marks of the state of vigilance are evident on all levels, from the cellular up to the level of the entire organ. However, for neurobiologists to assign reliably even a gross interpretation (e.g., loss of consciousness) to this activity, it is necessary to carry out measurements over time. Brain imaging shows that instantaneous activity patterns known to be evoked by visual stimuli can also arise spontaneously during anesthesia (Arieli et al., 1996; Kenet et al., 2003). Indeed state of the art EEG-based methods used for monitoring anesthesia during surgery (Gross et al., 2009) parcel measured signals into stretches of 2–30 s.[6] Thus, according to current neurophysiological understanding and medical practice, the identification and interpretation of brain states is inherently temporally extended. This suggests that, to the best of our knowledge, a brain "state" *is* a process extended in time, and so must be the activity-space patterns that constitute experience.

4.5. The problem of switching transient states

In an experimental paradigm referred to as backward masking, a visual stimulus is briefly presented to the subject, followed by a presentation of a second visual stimulus. Intriguingly, if the initial stimulus is presented for 50–100 ms, a duration that is quite sufficient for it to be consciously perceived, the onset of the second stimulus can nevertheless mask the first one (hence the name of the paradigm). The phenomenon of backward masking (along with the related forward masking) suggests that activity must be allowed to run its course for a certain time duration, if it is to give rise to conscious experience (Del Cul et al., 2007; Fisch et al., 2009).

For a system to give rise to experience, it must be active. If one insists that transient neural states be mapped to qualia, then by hypothesis the system is to be active while the quale stays fixed. This brings up the question of how neural states, and therefore qualia, are switched. Attaining a new state requires a finite period during which neural activity reorganizes. This process may extend to the entire brain (perhaps via some kind of global synchronization), or it may take place in local networks or circuits.

The first possibility does not seem to hold. For one thing, EEG data show that the state of alertness, which is associated with the richest experience, corresponds to less global synchrony (Steriade et al., 2001)[7]. Furthermore, if the degree of organization of the pattern of activity is associated with more phenomenal content, or richer experience, transitions between states would be associated with less content. This is certainly not the case.

Now if switching is a local phenomenon, it would preclude parsing the dynamics into global events due to overlap (multiple overlapping local transitions are equivalent to a globally smooth one). This would make any partitioning of activity into events arbitrary. Consequently, identification of transient states would be arbitrary, and with it the state to quale mapping.

4.6. The Dust problem

If experience unfolds frame by independent frame, and if a unique mapping exists between transient state vectors and fleeting experience, an evil demon could replace each frame in a given stretch, one at a time, with an arbitrary one, and we would never notice that they do not fit. What is at stake here is not the veridicality of experience as far as the world is concerned, but rather the correspondence between the experience and the physical processes giving rise to it. A broken correspondence renders some of the contents of experience nonsensical – for what could the feeling of surprise be if not the feeling of discrepancy between present experience and past experience? For experience not to be arbitrary, such surprise could only arise though the process of comparison. For a process to materialize, a system needs to go through a sequence of instantaneous states. Now, if each state in the sequence has meaning (content of experience, qualia) independent of the others, then reaching the same instantaneous state through different trajectories would nevertheless result in the same fleeting experience.

This issue is related to the so-called Dust Hypothesis, introduced and explored by the science fiction writer Greg Egan in his novel *Permutation City* (1995). Egan assumes that simulating someone's waking and active brain in a computer at a sufficiently detailed molecular level and temporal resolution generates a replica of that person's mind. He describes a series of scenarios in which such a simulation is subjected to progressively disruptive manipulation by an experimenter. In one of the experiments, only every 10th instantaneous state in the temporal order is computed; the replica or Copy, with whom the experimenter is in constant communication, notices nothing. In another experiment, the successive states are generated on spatially far-flung computers, with the same outcome.

Eventually, the plot thickens: the instantaneous states are generated in reverse temporal order, and the Copy is still oblivious to anything being different from normal. Egan (who must have read Searle) makes his protagonist conclude that, because temporal order does not matter, consciousness can arise, scattered like dust, in any medium that supports complex

[6] Note that *any* method would require measurements on the temporal scale of seconds. In particular, states of vigilance (and consciousness) are characterized by slow frequency components, e.g., delta waves (.2-4 Hz), whose detection necessitates measurements on the timescale of seconds. Furthermore, for a measure to be statistically robust, it must rely on a sufficient number of samples, suggesting that the commonly used 10 s. window is probably called for.

[7] Synchronization seems to be induced by global transition between up and down states. However, if synchrony is broken down into correlation in restricted frequency bands (what is known as coherency), consciousness is associated with greater local synchrony (Fekete, Pitowsky, et al., 2009; Fisch et al., 2009).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

instantaneous states. Chances are that once in a long while a random state would be isomorphic to that of some mind, which would thus assemble itself out of the dust.

We tend to interpret this story as a reductio ad absurdum, its upshot being that a model in which qualia are equated to instantaneous independent states would be hard pressed to explain various mundane cognitive phenomena, not the least of which is the experience of time. We shall return to this troubling issue in Section 5.[8]

4.7. Physiological evidence

In recent years, direct physiological evidence has been accumulating against the idea that experience can be explained or modeled by mapping instantaneous states to qualia. The brain mechanisms whose activity is implicated in giving rise to experience are extended in time over and above the trivial sense in which every physical system is (through obeying dynamical equations that express the laws of physics). We briefly sketch out several examples of the relevant evidence.

First, intracellular recordings show (Azouz & Gray, 1999, 2000, 2003) that the firing threshold of neurons is constantly adjusted to reflect the statistics of the membrane fluctuation in the preceding 200 ms or so (and not only to reflect the mean value, but also the slope of the change). In a similar vein, in a recent study using both human and monkey subjects, neuronal data were collected using electrode arrays implanted in several motor areas (Truccolo et al., 2010). The instantaneous spiking of the recorded neurons was predicted using one of three models, listed below from the least efficient to the most efficient: (1) an Ising model (i.e., predicting the activity of single neurons from the instantaneous state of the recorded ensemble in the preceding time step), (2) the spiking history of the neuron in the preceding 100 ms; (3) the spiking history of the ensemble in the preceding 100 ms. Finally, several modeling studies suggest that there are other electrochemical cascades apart from neuronal firing, that are nevertheless cognitive or representational, yet operate at much longer time scales. For example, a recent study (Mongillo et al., 2008) showed that memory traces can be retained by calcium dynamics (rather than persistent spiking; e.g., Amit, 1995), memory being a medium for context and expectations, which are both essential ingredients in the making of experience.

5. A sketch of a viable computational theory of experience

5.1. Dynamics to the rescue: state-space trajectories as representational primitives

The previous section showed that explaining experience in terms of a state to quale mapping is doomed to fail. However, at several points of the discussion a positive alternative seemed to emerge – namely, construing activity (as well as experience) as a dynamic process eliminates some of the quandaries that beset theories based on instantaneous states (cf., Spivey, 2007).[9] We now develop this line of argument by describing how construing activity as trajectories in the system's instantaneous state space (Churchland, Cunningham, et al., 2010; Churchland, Yu, et al., 2010; Churchland et al., 2007; Geffen et al., 2009; Mazor & Laurent, 2005; Yu et al., 2009)[10] is a natural starting point for formulating a computational theory of experience.

First, as we noted above, experiential state dependent activity is an inherently temporal phenomenon. Hence, state-space trajectories readily support the realization of representational (experiential) state, simply by extending in time. Given the experimental data, trajectory segments that realize distinct and well-defined content should extend anywhere from hundreds of milliseconds (as suggested by masking experiments) to seconds (the duration of the window used by current measures of depth of anesthesia).

Second, by sweeping through the system's instantaneous state space, trajectories blow away Egan's Dust trope. As expressions of content, they can naturally represent content associated with processes, not the least of which is the experience of time itself.[11] They make possible the realization of experience that is rooted in diachronic comparison, as well as of content that is fundamentally dynamic (e.g., sound and movement).

Finally, positing trajectories as expressions of phenomenal content constitutes a significant step towards embodying the modal structure needed to realize experience within the confines of activity. The reason is that a trajectory, being an orderly succession of configurations, places much-needed significant constraints on the causal mechanisms that could result in the given path through the system's instantaneous state space, given the initial conditions (i.e., the starting instantaneous state).

While trajectories do narrow down the range of possible mechanisms that could give rise to them, they fall far short of singling out a unique mechanism. However, as we shall see in Section 5.4, defining trajectories as the physical isomorphs (realizers) of phenomenal content provides the foundation needed to complete the explanatory move.

[8] Chalmers's refutation of Searle's charge that the computational theory of mind is tantamount to panpsychism hinges on the states of the system that implements a mind being in the right temporal order (Chalmers, 1994).

[9] Spivey (2007, p.305) writes: "... you might be tempted to imagine your mind as a kind of floating ball that moves around in the high-dimensional neural state space of the brain. What you have to be careful about, however, is conceiving of your self as the equivalent of a little homunculus sitting on that ball going along for the ride. You are not a little homunculus. You are not even the ball. You are the trajectory."

[10] These studies lend some support to the notion that the pertinent way of formulating the dynamics of at least some neuronal systems (olfactory, motor and auditory) is through a trajectory in a low dimensional space.

[11] This raises the question whether computational models of experience need to be constrained such that they would exhibit time-asymmetry.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

5.2. Self-interpretation

As we insist throughout the present article, the central guiding principle for a theory of experience must be that it should offer an account of experience intrinsic to the system in question. Unlike a time-frozen snapshot of a system's instantaneous state, which needs an external interpreter to qualify as a quale, or, indeed, a representation of anything, the dynamical approach outlined here conforms to this principle.

To see this, we note that the word "representation" is used in two fundamentally different ways in the cognitive sciences. The first of these has to do with *potential* interpretation. A written sentence, for instance, is a representation only to very particular systems, namely, those equipped with the right kind of sensory apparatus and background knowledge (e.g., of English). Even for a properly equipped system, however, the representation remains potential until it undergoes an interpretative process – and any such process by definition takes time. It is this interpretative process that fixes the second sense of representation, and it is on this latter sense that we focus on in the present work.

Now, the trajectory of a dynamical system through its state space serves as its own interpretation. A dynamical system that is allowed to run its course (merely by following the laws of physics; cf. Section 2) interprets its own instantaneous states in the following sense: each instantaneous state, along with its history (to the extent and in the form mandated by the system's dynamics), determines the next state, in a manner that is properly counterfactually validated (again, by the system's dynamics).

If experience is indeed inherent to the system's dynamics, the question "how does the system know what experience it is having?" receives a naturalistic answer that avoids positing an external agent or an internal homunculus charged with interpreting its states. In particular, if the system undergoes two distinct experiences (say, evoked by two distinct stimuli), the only necessary interpretative process is for the system to respond to the two differentially (i.e., by two distinct trajectories), subject to the constraints that we already stated – (i) that each trajectory be endowed with the appropriate intrinsic structure to embody the experience it does (i.e., be its physical isomorph[12]), and (ii) that it occupy the right locus in the space of all possible trajectories (in terms of belonging to the pertinent clusters in the realized conceptual domain). Simply put, for a truly intrinsically representational system, trajectories *are* interpretations. Experience, *qua* discrimination between the two kinds of stimuli, is thus intrinsic to the system's dynamics.

5.3. Categorization capacity, learnability, and VC dimension

We shall now offer a possible formalization of the idea of graded qualia, or Q-cline, following Edelman (2009). For concreteness, the following discussion is placed in the context of vision. To quantify the experiential power of a representational system that treats visual scenes as points in some measurement space, $s \in S$, consider how it can distinguish the scenes from one another by classifying them with respect to a set of concepts C. Think of it as making a set of distinctions (flagging objects, textures and the relations between them, and so on, all of which are members of C) that express the structure of the current experience. Jointly, all these distinctions also form a concept ("a boy running in the street, lined by cars, on a rainy day . . ."[13]) in C. A system is capable of having complex experiences to the extent that it has both a high-resolution measurement front end and a rich conceptual back end (a multi-megapixel digital camera ca. 2011 and a blind person both fail to see, for complementary reasons). If the dimensionality of the measurement space is sufficiently high, the system in question will be able at least to encode a very large variety of distinct scenes. Let us, therefore, assume that the dimensionality of the measurement space is high enough and proceed to examine the role of *conceptual richness* in seeing.

This can be done by formalizing the representational system's conceptual back end as a classification model. The model's power can then be expressed in terms of its Vapnik–Chervonenkis or VC dimension (Vapnik, 1995; Vapnik & Chervonenkis, 1971). Consider a class of binary concepts $f \in C$ defined over a class of inputs (that is, measurements performed over scenes from S), such that $f{:}S \rightarrow \{0, 1\}$. The VC dimension $VCdim(C)$ of the class of concepts (that is, of the model that constitutes the categorization back end of the visual system) quantifies its ability to distinguish among potentially different inputs (similar schemes for multiple-way and fuzzy classification exist).

It is natural to think of discrimination geometrically, in terms of objects represented by points in a feature space. A class of concepts corresponds to a class of decision surfaces; in a 2D plane, for instance, these could be straight lines or circles or parabolas. A more complex class affords a finer discernment ability; thus, parabolas in 2D (three-parameter "surfaces") can be used to single out more configurations of points than straight lines (which have two parameters).

A concept class is said to *shatter* a cloud of points representing a set of n items in some feature space if its member concepts can support all 2^n ways of dividing those points into two complementary categories. For example, the class of straight lines in a plane shatters any set of 3 points, but not any set of 4 points. This leads to the definition of the *VCdim* of a concept class C as the cardinality of the largest set of inputs that a member concept can shatter.

[12] Following Shepard (1987), it is common to distinguish between direct isomorphism, and second order isomorphism (Edelman, 1999; Shepard, 1987). However to say that activity corresponds in structure to the experience it realizes, is to say that there are measures in place which can quantify certain aspects of an instance of activity. Thus, just like in second order isomorphism, the isomorphism here would be mapping of essential structure (i.e. functions) between the domains (Fekete, 2010).

[13] Of course a comprehensive listing of the content of even a fleeting experience is much more complex than that (and messy!), so we ask for the reader's indulgence regarding the shorthand.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Fig. 1. A pattern is constituted not only by "filled out" areas, but by the "empty" areas as well – take these or the others away and you end up either with a blank sheet or an undifferentiated "blob." Similarly, detection of patterns in a high dimensional space is enabled by detecting "holes," or regions of low density (see Fig. 4).

Because classifying a scene as being an instance of a concept amounts to seeing it *as* something, we have thus effectively formalized the notion of "seeing as" introduced by Wittgenstein (1953; cf. Edelman, 2009). We are now ready to extend this framework to encompass "raw" phenomenal experience, or the ability to "just see." The key observation is this: among several conceptual systems that happen to share the same measurement space, the one with the highest VC dimension is the most capable of distinguishing various subtle aspects of a given input. In other words, to progressively more complex or higher-*VCdim* visual systems, the same scene would appear richer and more detailed – a quality that translates into the intuitive notion of a progressively better ability to experience the scene, or "just see" it.[14]

By formalizing and quantifying the richness of conceptual representation systems, the *VCdim* framework that we just outlined contributes to the construction of a viable computational theory of experience. However, while the *VCdim* framework addresses issues of representational capacity, it does so without rooting capacity in the intrinsic structure of activity trajectory spaces. In the next section, we take up this crucial task.

5.4. Representational capacity

We now recast the notion of representational (experiential) state in terms that are appropriate to the conception of activity as trajectories through a state space (Fekete et al., 2009). What is it that makes activity trajectories fit to express experiential content? For example, what are the systematic structural changes in activity that correspond to going from dreamless sleep all the way to full wakefulness?

If systematic change in the richness of experience corresponds to a change in representational (experiential) state, the richness of experience remains constant when the representational (experiential) state is fixed. We can say then that given a representational (experiential) state, the complexity of experience is invariant, and so must be the complexity of activity trajectories. What happens when the experiential state changes?

As one emerges from the oblivion of dreamless sleep, one is able to take in more and more details of the surroundings. To do so, the system must be able to make finer and finer discernments regarding both the internal and external environment. A change in representational (experiential) state is thus associated with change in the conceptual structure realized by activity trajectories. At the same time, as experience becomes richer, and with it the realized conceptual domain, the structure of activity trajectory *space*, which encompasses all trajectories that are possible under the current regime, should become more complex to accommodate this. As noted above, this should result in the formation of increasingly complex structures of clusters in activity trajectory space.

If richer experience necessitates more complex activity trajectories, as well as increasingly complex structures of clusters in the space of activity trajectories, these two facets of the complexity of activity must be coupled: the subtler the discernments (differentiation in the phenomenal field) that arise from the representation of one's surroundings, or mental content in general – which is manifested as enhanced clustering in trajectory space – the richer the experience, and consequently the complexity of activity trajectories. But the converse must be true as well: as activity trajectories grow more complex, so must experience, and with the richness of experience the distinctions that are immanent in it, and hence the complexity of the

[14] It is worth recalling that the VC dimension of a class of visual concepts determines its learnability: the larger *VCdim*(*C*), the more training examples are needed to reduce the error in generalizing *C* to new instances below a given level (Blumer, Ehrenfeucht, et al., 1986; Edelman, 1993). Because in real-life situations training data are always at a premium (Edelman, 2002), and because high-*VCdim* classifiers are too flexible and are therefore prone to overfitting (Baum & Haussler, 1989; Geman, Bienenstock, et al., 1992), it is necessary to break down the classification task into elements by relegating them to purposive visual sub-systems. Being dedicated to a particular task (such as face recognition), such systems can afford to employ the simplest possible classifier that is up to the job. For this very reason, simple organisms whose visual system is purposive and therefore good at learning from examples are poor general experiencers. The rich and purposeless experience of "just seeing" means being able to see the world under as many as possible of its different aspects, an ability which corresponds to having a high *VCdim*.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

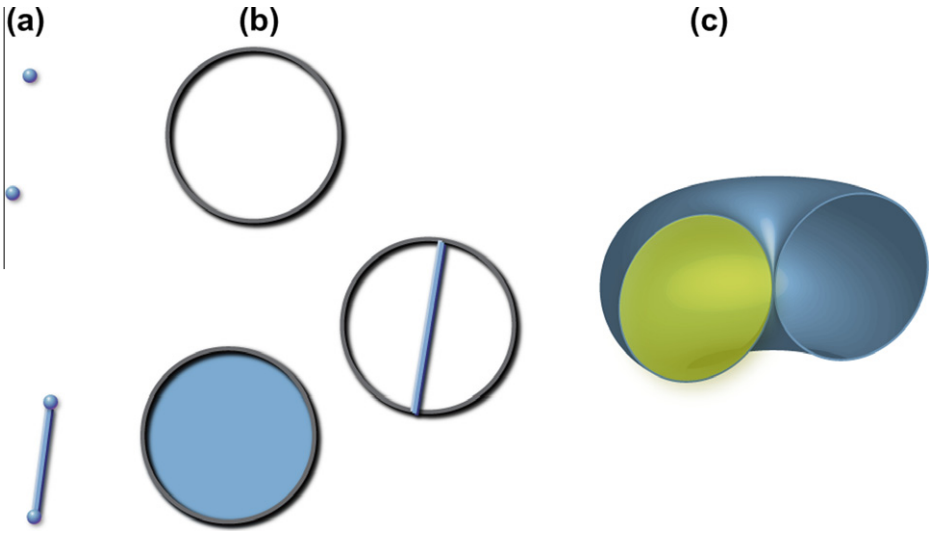

Fig. 2. Configurations and hence holes come in all dimensions: (a) two disconnected points (top) are separated by a 1-D hole – it can be filled out by a line (bottom), (b) a ring structure (top) comprises a 2-D hole – it cannot be filled out by a line, which would only separate the hole into two ones (middle). Only a 2-D surface (bottom) can fill out the hole. Similarly in higher dimensions: in (c) it can be seen that a surface is not sufficient to fill out a 3-D hole (cavity), as seen by looking at the cross section of the torus.

realized conceptual domains. We therefore define the *representational capacity* of a space of trajectories as the joint (tightly coupled) complexity of (i) the structure of individual trajectories in it and (ii) the structure of the space of trajectories itself.

To move from these general considerations to operational terms, let us first consider how the complexity of the structure of a space (such as a space of trajectories), that is, configurations of clusters, can be measured. As noted above, a reasonable measure of complexity will be sensitive not only to the degree of clustering found within a space, but also to the effective dimensionality of the various configurations of clusters to be found within that space. So in essence what we would like to be able to do is simply count configurations of clusters according to their effective dimensionality.

How would one go about this computationally? Imagine a pattern of ink drawn on a sheet of paper (Fig. 1). Our usual way of thinking would be to conceive this pattern as being formed by the regions of the sheet covered with ink. But in fact the empty regions are just as indispensible in making up the pattern – were the paper covered entirely in ink, no pattern would emerge, just as if the paper were left blank. If instead of paper and ink we think of a high dimensional space and clusters of points in that space, we see that blank regions in space, or holes, are the mark of configurations of clusters.[15] The latter observation holds the key to our problem – counting holes is a relatively straightforward affair.

To realize this, note that a hole in space is in fact a region in which the density of points is zero. So measuring the density of points – let us say by chopping up space into regions[16] – would enable us to register holes in space and thus configurations. To complete the picture, we should note that holes vary in their dimensionality (see Fig. 2). For example, a ring configuration could be turned by a 2-dimensional "patch" into a disk, and hence is characterized by a 2-D hole (which is referred to as a *1-hole* for technical reasons). Similarly, a sphere (such as a basketball) is formed by a 3-dimensional cavity (a 2-hole). Thus, racking up holes according to dimension enables enumerating configurations of clusters according to their effective dimensionality.

There is a twist to this, however, as a configuration is essentially a function of scale or resolution. To realize this, imagine a configuration of clusters comprised of configurations of clusters (see Fig. 3). If we look at the figure from a bird's eye view (bottom right), which would be equivalent to looking at the picture at a lower resolution or larger scale, only the gross level organization becomes apparent. To see that each cluster is indeed itself a configuration, we would have to zoom in (top left), that is, increase resolution or decrease scale. However, given the limited range of scales that can be considered simultaneously, this would mean that one would lose the big picture, i.e., the higher level configuration. It turns out that exactly this information, namely, the number of configurations of clusters according to dimension as a function of scale, is readily computable by the *multi-scale homology* of a space and can be represented as *Betti graphs* (see Fig. 4 for an illustration and Fekete et al. (2009) for technical details).

In comparison to clusters, measuring the complexity of trajectories is a much more straightforward affair. Recall that our considerations led us to realize that the complexity of activity trajectories is an invariant, given a representational (experiential) state. Available evidence suggests that suitable invariants may have to do with the spatiotemporal organization of

[15] To forestall an outcry from readers familiar with homology theory, we note that while some patterns – such as the letter S – strictly speaking do not contain holes, once the technicalities of measuring homology from point clouds is taken into account, a different picture emerges. In such scenarios, homology *must* be measured as a function of scale, and thus at certain scales an S shape will have the topology of two rings joined by a point.

[16] In fact there are much more elegant and parsimonious methods, which face up to high dimensionality markedly better than the naïve approach (de Silva & Carlsson, 2004; Edelsbrunner, Letscher, et al., 2002).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

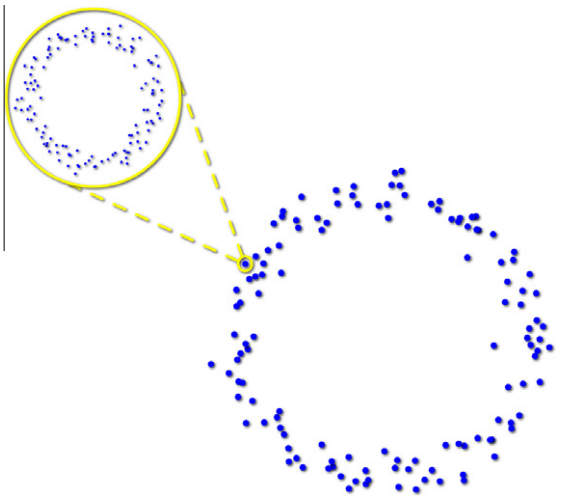

Fig. 3. *Bottom right*: a "bird's eye" view of a cluster in a representation space. *Upper left*: a closer look at one of the elements of the coarse-scale cluster reveals that it possesses structure on the finer scale. See text for details and discussion.

activity (Cao et al., 2007; Contreras & Llinas, 2001; Leznik et al., 2002; Makarenko et al., 1997). For example, it has recently been shown that various complexity measures of neuronal trajectories increased monotonically with wakefulness (Fekete et al., 2009). In other words, activity trajectories can be classified according to representational (experiential) state: a classifying function, which we will refer to as a *state indicator function*, can be defined on activity trajectories (i.e., over the space of activity trajectories). A state indicator function assigns each trajectory a number[17] so that a given state of vigilance (or consciousness) is associated with a typical or *characteristic value*.

This brings us to the crux of the matter: if constructed properly, a state indicator function provides a means for measuring representational capacity. As just noted, the characteristic value of a state indicator function would pick out all activity trajectories in a given representational (experiential) state, as ex hypothesi they share the same degree of complexity. In other words, it would single out the entire subspace of activity trajectories associated with a representational (experiential) state. In technical terms, this amounts to saying that the *level sets*[18] of a state indicator function carve out experiential state-dependent spaces of activity trajectories. As these are well defined mathematical objects, their complexity, as measured by their multi-scale homology, can be computed exactly. In other words, a state indicator function provides a handle on the otherwise elusive concept of the space of all possible trajectories, and therefore on the space of possible experiences (see Appendix A for additional details).

The notion of a state indicator function offers another perspective on the definition of representational capacity. Such a function can be thought as expressing a list of constraints. An activity trajectory must satisfy these constraints to be part of an activity trajectory space realized in a given representational (experiential) state. If these constraints are lax to the degree that any trajectory meets them, the resulting space would simply be homogeneous and structureless. The more stringent they become, fewer and fewer trajectories are able to satisfy them, which will eventually lead to the formation of holes in the space of viable trajectories. As holes are the mark of configurations of clusters, we see that if a system constrains its dynamics in the right way, this will lead to a complex cluster structure in the realized space of activity trajectories.

These ideas were applied to primate neural imaging data (Fekete et al., 2009). First, a state indicator function, which could perfectly classify activity trajectories according to state of vigilance by their complexity, was constructed. It was then found that the complexity of the level sets associated with characteristic values of this function increased with wakefulness.

The notion of representational capacity explicates how the aspects of activity needed to realize experience are immanent in the causal structure of a system: if a system constrains the dynamics in a certain way, a non-trivially structured space of activity trajectories emerges. But if activity trajectories at a given experiential state exhibit invariant spatiotemporal properties (complexity), this means both that (1) the trajectories that express experiential content are complex in exactly the right manner to realize the phenomenal content that they do, and (2) that this complexity is intrinsic because it inseparable from the complex structure of clusters in the space of activity trajectories. This explains how changes in the richness of phenomenal experience go hand in hand with the system's ability to make distinctions: realizing a complex quale is physically inseparable from realizing an entire rich *space* of qualia. In other words, this approach expresses the potential structure needed to realize a conceptual domain solely in terms of the intrinsic and actual makeup of activity trajectories.

[17] Or a low dimensional vector; cf. (Hobson, Pace-Schott, et al., 2000).

[18] For a state indicator function $SIF : A \to \mathscr{R}$, the level set associated with a value $c \in SIF(A)$ is the entire set of activity trajectories $a \in A$ for which $SIF(a) = c$, or all the activity trajectories that a state indicator function would assign the same score (value) – that is, they would exhibit the same degree of complexity (as measured by the SIF).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

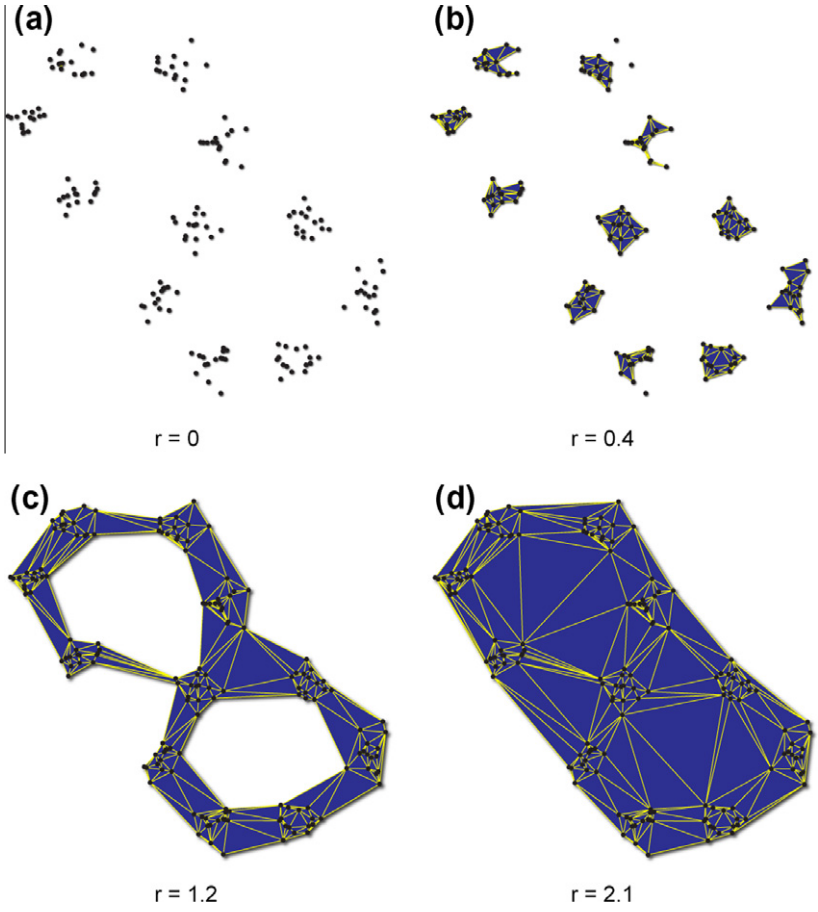

Fig. 4. Counting configurations of clusters via multi-scale homology. In the idealized scenario above a joined two ring structure (an 8) is sampled at regular intervals with measurement noise. This results in several clusters scribing the general layout of the underlying shape (i.e. space). In order to glean the underlying structure, the points in the cloud need to be "merged" to form a shape. This is done by joining some points with lines, and filling out some of the formed triangles with surfaces (and similarly in higher dimensions). The resulting construct is a *simplicial complex*. According to which points are connected with lines (and filled out with surfaces) a very different picture emerges – as can be seen in the figure above. Hence the objective way to go about it is to use a metric criterion – i.e. only connect points if they are "close enough". However, as closeness is in the eye of the beholder, what is done is to simply create all possible complexes as a function of scale (metric criterion – in the example below r denotes the Euclidean distance between points): The complex in (a) results from using $r = 0$, i.e. connect with a line only points which are closer than 0. Hence none of the points connect. In (b) $r = .4$, which picks out the underlying clusters (which in our case is simply the variance of the noise). In (c) $r = 1.2$, in which case the underlying structure is readily apparent. Accordingly by purely algebraic means this complex allows one to compute that number of holes present in the complex is 2. These are actually *2-D* holes – they can be filled out with a surface. Increasing r (cf. 2.1 in (d)) will eventually result in a blob lacking structure. The resulting information – the number of holes of a given dimension (in this case 2-D) as a function of scale – is represented by the *Betti profile*.

Our definition of representational capacity can also contribute to solving the problem of silent units. The key observation here is that change in the fraction of silent units changes the complexity of trajectories: if all units are active (as in a seizure), activity loses structure; ditto if all units are silent. More generally, the spatiotemporal organization of active units is what defines the complexity of trajectories, which could not materialize without the complementary spatiotemporal profile of silent units. As it is exactly the characteristic complexity of activity trajectories that shapes the structure of activity trajectory space, we see that silent units are, as it were, more of a solution than a problem.

So far we have discussed representational capacity in the context of measuring the experiential state of a given system, showing how it can serve to quantify the richness of experience in a given experiential state, as well as establishing when consciousness is lost. Note, however, that a complexity measure also establishes an ordering over the space of systems by their representational capacity, thereby also ruling out some classes of systems as non-conscious. To reiterate, systems that give rise to homogeneous (topologically simple) activity trajectory spaces lack consciousness altogether. That said, it is important to stress that by no means are we implying that the structure of a trajectory space alone suffices to realize experience. Rather, only activity trajectory spaces that are parceled into non-trivial level sets by a plausible complexity measure fit the bill. For the interested reader, we offer in Appendix B an analysis of two examples of dynamical systems that give rise to non-trivial trajectory spaces, yet fail to meet the criteria articulated herein.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

5.5. Facing up to the curse of dimensionality – multiscale organization

We have seen that computational analysis of experience needs to be carried out at the level of the entire system, that is, in terms of the correspondence between the space of activity trajectories and the conceptual/perceptual domain that the system realizes. It would seem that the dimensionality of the spaces involved, especially in complex systems such as the human brain, may preclude any such attempt from succeeding. To show that this need not be the case, we revisit our formulation of representational capacity.

In operational terms, representational capacity – the coupling between the complexity of activity trajectories and the complexity of the trajectory space – was defined as the multi-scale homology (geometric complexity) of the level sets of a state indicator function (experiential state-dependent space of possible activity trajectories). This definition is very general and is applicable to any dynamical system, provided that proper measures of complexity are used. Now, as we noted before, in complex organisms the mark of experiential state is apparent in brain activity, from the cellular level all the way to the entire organ. Accordingly, were we to confine our observations to parts of the system, the same game could be played, i.e., we could define state indicator functions, which would classify activity trajectories according to (local) experiential state and single out the experiential state-dependent activity spaces via their characteristic values, and hence level sets.

Of course, the resulting activity trajectory spaces could no longer be analyzed in terms of the conceptual domain realized by the entire system: we could not infer the entire structure from its parts without knowing how the respective partial activities behave jointly. The potential benefit of the subsystem-based approach is that makes it possible to manage the dimensionality of the entire system. The down side is that a system that relies on this strategy to keep its dynamics under control would have to coordinate the activity of its sub-systems, in order to rein in the complexity of the joint trajectory space.

This brings us back to the idea of coarse-graining, introduced before: we propose that representational systems realize conceptual domains by implementing the same computational scheme at several levels of organization at the same time. Constraining the complexity of dynamics at various levels of organization gives structure to representational trajectories, while at the same time giving rise to an activity trajectory space with the necessary complexity of structure to support the goals of the system. This strategy thus allows the formation of complex hierarchical conceptual structure, while keeping the dimensionality of the solution manageable.

As we have seen, a hierarchical conceptual structure is realized by a hierarchical structure of clusters. If the system is to utilize this hierarchy and operate on it (or modify it through learning), it needs to be attuned to the realized structure of clusters at several scales (in the metric sense). Again, a straightforward way of doing so is to allocate distinct systems to deal with the various levels of the structure (i.e., concepts of various degrees of abstraction, in accordance to their level in the hierarchy of clusters). From this perspective, it is clear why in the brain different brain areas seem to be dedicated to representing content at various levels of abstraction (for a more comprehensive analysis, see Fekete, 2010).

It may seem that partitioning a given system (which, moreover, can be carried out recursively) is arbitrary, and therefore that countless distinct, yet seemingly equivalent, computational models can arise, all of which would nevertheless be equally applicable to the system. As we noted above, such arbitrariness, if true, would cast doubts on a computational theory of experience that adopts this approach. Happily, arbitrariness of hierarchical structure is not an issue in real nervous systems, which exhibit various readily discernible levels of histological and anatomical organization (e.g., intracellular apparatus, dendritic and axonal architecture, cells, nuclei, mini-columns, Brodmann areas, tracts, and hemispheres). Our analysis can thus be seen as including the hypothesis that the hierarchical organization of brains constitutes the requisite well-defined natural way of partitioning the computational architecture through which the brain realizes experience.

In summary, we have seen that by resorting to coarse-graining, a system can not only attain and manage a complex hierarchical representation, but also bring under control the vast dimensionality of its activity trajectory space, by essentially implementing the same computational scheme at different levels of granularity.

5.6. Carving out conceptual domains

As we have shown, for a system to give rise to consciousness, it must realize a complex and highly structured space of activity trajectories, and thereby instantiate a conceptual domain. However, complex organisms do not set out to achieve some preordained conceptual domain, but rather *a* conceptual domain, namely, one that suffices to further the organism's interests (subject to quite extensive hardwired constraints). On the one hand, it is advantageous for the representational system to use its experience in adapting the structure of the conceptual domain it realizes to better reflect the organism's surroundings; on the other hand, doing so at too high a level of detail or fidelity may be too expensive or simply infeasible. What is feasible, though, is to apply representational machinery that funnels the dynamics of the system during maturation to some normative level of complexity (characteristic spatio-temporal structure) throughout the various sub-systems it comprises, and in the process of doing so establishing a complex hierarchically clustered space of trajectories.

To carve up a complex hierarchical structure of clusters in the representational medium, the system's learning mechanisms must be capable of elaborating the structure of the resulting trajectory space, so that new clusters will emerge and existing configurations will be reworked so as to increase the salience of certain features of the structure of the trajectory space. Unsurprisingly, this fits the intuitive notion of learning: as we learn something new, we come to be able to draw

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

new distinctions (which corresponds to the emergence of new clusters) and/or hone existing ones (which amounts to making a structure of clusters more pronounced in the metric sense).

Naturally, not any old elaboration will do; rather, the emerging structure needs to reflect the structure of the world (modulo the particular interests and goals of the agent). A useful way of posing this problem rests on the observation that as far as the system is concerned, the world actually is a complex multi-dimensional distribution of sense data resulting from the various interactions of the body with its environment mediated through various receptor arrays (Mach, 1886).

The most straightforward solution to this problem would be to simply strive to distill the structure of the world distribution, so that the trajectory space more or less mirrors it. However, it is clear that this is not feasible, as the world offers much more information than an animal could ever learn given the means and the time at its disposal. Accordingly, a system must prioritize its information extraction and representation. This would amount to extracting some features of the world distribution, while remaining oblivious to other such features. For example, after studying music, one's experience even of pieces that were known before is markedly altered. One can notice and appreciate structure that was there all along (nothing changed in terms of the activations of auditory receptors). To summarize, a learning mechanism needs to carve out the space of activity trajectories, so that it will pick up and enhance structural aspects of the world distribution, while overlooking and downplaying other such features.

In the process of learning, a system would have to reconfigure itself. For example, in neuronal systems, it seems that a major mechanism enabling learning is the adjustment of the strength of the synaptic connections between neurons. However, doing so will likely have an effect on the resulting dynamics as learning accumulates. Thus, learning has to be regulated in order to result in increasingly complex dynamics (at least during maturation; for mature organisms it might be just a tugging of the blanket affair), to achieve the goal of increasing the system's representational capacity by keeping the complexity of activity trajectories coupled with the complexity of the space of trajectories.

The upshot of all this is that a full-fledged theory of phenomenal experience will need to explicate several computational mechanisms. Among these are learning mechanisms that could cause trajectory space to differentiate into clusters, and the mechanisms through which dynamical systems could constrain their dynamics to follow a plausible complexity measure. Until proven otherwise, it makes sense to assume that such mechanisms will have a universal form – at the least where brains are considered. Therefore, we can summarize by saying that a theory of experience would involve articulating the fundamental equations (laws) of complexity, as well as the fundamental equations (laws) of learning.

5.7. What next?

Before concluding, we would like to discuss in broad outline the kind of theoretical and empirical research that would help bring the ideas presented here to fruition. The working hypothesis for research outlined below is that neuronal machinery exercises tight control over the dynamics it gives rise to, such that: (1) activity at any given time results from a strict regime (representational state), which belongs to a spectrum of such regimes, and in which the spatiotemporal makeup of activity trajectories is invariant under a complexity measure; (2) normal representational states (e.g. sleep, wakefulness) form families of stable regimes that the system can enforce, and switch between, according to circumstances; (3) the system molds the cluster structure of the experiential state dependent activity trajectory spaces to serve its interests.

Data collection. The first order of business here is to characterize representational states in some detail. To that end, one must collect physiological data that would allow exact articulation of the spatiotemporal structure held invariant in different states of consciousness. This would entail collection of ongoing activity data in normal alert states, as well as in sleep, during dreaming, in states that are altered by the influence of various psychotropic agents, etc.[19] While some work along these lines has been done in recent years (e.g., Kenet et al., 2003), neuronal data are still by and large collected in artificial, tightly controlled stimulus–response scenarios.

Understanding how representational states are maintained and switched. Switching is probably effected in part by global mechanisms, such as neuromodulation and "pacemaker" circuits. However, representational states are also shaped by local mechanisms such as homeostatic processes at the single neuron level that promote invariance in cell spiking activity over time (e.g., Rabinowitch & Segev, 2008; Azouz & Gray, 1999, 2000, 2003).

Elucidating the relation between learning and state-dependent activity. Developmental studies aimed explicitly at quantifying the changes in state-dependent activity brought about by maturation and learning will shed light on the "global" mechanisms at the heart of learning, namely those that result in greater differentiation into clusters in trajectory spaces.

Studying representational state maintenance and switching in artificial neuronal networks. Apart from current lack of data and insufficient empirical generalizations (which are after all the substrate of modeling), another source of concern is that the work being done along these lines usually targets isolated characteristics of the circuit (e.g. phase coherence El Boustani & Destexhe, 2009; or, exhibiting up and down states Holcman & Tsodyks, 2006). It remains unclear to what extent such results generalize to the relevant scenario — single networks that can simultaneously express and maintain many such characteristics.

[19] Recording of activity both in the absence and in the presence of stimulation will have to be considered, the latter focusing on rich natural settings (e.g., movies, natural speech), which can serve to stabilize the experiential states of subjects (c.f., Hasson, Nir, et al., 2004).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Further development of the theory of learning in high dimensional spaces. This necessitates a transition from mass univariate statistical procedure, to multidimensional hierarchical schemes that will allow one to utilize the volume of data afforded by measuring population activity to offset the inherent limitations associated with high dimensionality. One possible way to do that is through clever sampling schemes geared towards approximating high dimensional structures, such as trajectory spaces (e.g., Bendich et al., 2010; Bubenik & Kim, 2006; Hudson et al., 2009).

6. Conclusions

Unless we got it completely wrong, the research program following the path we have sketched above could culminate in computational models in which each and every property of experience would be given a concrete expression in the form of explicit, tractable equations and measures, which would articulate in the language of mathematics the correspondence between a system's experience and its activity. This would make it possible not only to understand and manipulate conscious experience in exact and specific ways, but perhaps also to construct artificial[20] conscious systems. Will this then amount to closing the infamous explanatory gap?

On the one hand, the case could be made that if this were to come to pass, consciousness will have been elevated to the lofty summits inhabited by theories we hold to be paradigmatic of scientific explanation. Others might be left disappointed, though, arguing that the explanatory canvas painted by such a theory would be just as complete were we all zombies. While we sympathize with the latter concern to some extent, we would like to emphasize that consciousness is a fundamental property of reality: not only is it a fact, but only through it can we derive other (therefore secondary) facts. Just as science in general cannot explain the *existence* of fundamental properties of reality (such as gravity), yet can go a long way toward explaining reality once those fundamentals have been cast in the form of "laws" (i.e., equations), so a mathematical theory of consciousness that satisfies the viability criteria stated in this paper can go a long way toward consolidating the science of the mind.

In a similar vein, it is often argued that explaining consciousness requires specifying its function, lest it remain epiphenomenal. Seeing that we have been quite explicit with regard to what consciousness is, we can offer a naturalistic answer to this question, namely, that the function of consciousness is world-building. In other words, consciousness distills aspects of the statistics of the environment as captured by the senses into a complex hierarchical space of activity trajectories, or, equivalently, organizes the information impinging on the body into a coherent whole – a conceptual domain.

In fact, our analysis puts us in a position to offer by way of hypothesis a much more satisfying, albeit less precise, answer: the function of consciousness is to enable thought and the exploration of ideas (i.e., conceptual and perceptual domains). The reasoning is straightforward: for concepts to be manipulated, they must first be represented or realized, and as we have argued, it is exactly this which gives rise to consciousness in the first place.

Acknowledgments

We thank Kat Agres, David Chalmers, Axel Cleeremans, Rick Dale, Barbara Finlay, Rafael Malach, Björn Merker, Thomas Metzinger, Helene Porte, Michael Spivey, and two anonymous reviewers for their comments on a draft of this paper.

Appendix A. State indicator functions

In what follows, we briefly outline one possible way of realizing a state indicator function. At this point, seeing that pertinent theory – the theory of complexity, the theory of learning in high dimensional spaces and the theory of neuronal systems – is seriously underdeveloped, our construction will be geared toward data mining and gross phenomenology.

To recapitulate what was mentioned in the body of the paper, a state indicator function associates with each activity trajectory a score, which varies according to representational state. The state indicator function encapsulates within it the essential characteristics of the dynamics of a system – i.e., while it is defined on a high dimensional space at large, it associates disparate scores with trajectories that could be produced by system and with those that could not.

The crux of the matter is that a level set of this function encompasses all the possible activity trajectories corresponding to a representational state, and is therefore in effect a model of the system's state dependant trajectory spaces. Hence, to the degree that it is appropriate, it can allow one to reconstruct the underlying spaces given a reasonable sample of the systems state dependant dynamics. The idea is to construct SIFs, while constraining them to have certain desired properties, to ensure that the ensuing models of the state dependant trajectory spaces will exhibit desired properties. For example, it seems that it is of great import that the SIF function be smooth – as this causes the level sets of the function to be smooth manifolds (for other benefits of smooth representation see Edelman, 1999). This ensures that the resulting spaces will be metrizable, which corresponds with a fundamental property of experience – namely, that it is organized according to the similarity of the contents of experience. Another way of putting it is that any given stretch of experience could have been infinitesimally different along any aspect, content, or property it comprises. The mathematical equivalent is captured by the manifold property.

[20] Note that simulations of neuronal networks on digital machines violate many of the necessary preconditions for experience, spelled out above (see especially Section 4 and Appendix B), hence if at all possible, machine consciousness must be of a different kind.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

A state indicator function can be constructed in parts, to achieve any degree of specification over the properties of the trajectories it will sanction as belonging to a given representational state. First, various measurements pertaining to complexity can be carried out on activity trajectories – such as various measures of randomness, spatio-temporal organization, and correlation. Next, a classifier function (e.g., a nonlinear Fisher discriminant) can be fitted to these measurements (i.e., defined over this feature space). The classifying function can then be further elaborated to encapsulate any measurable property of activity trajectories, which might not be conducive to dissociation between representational states, but are nevertheless characteristic of neuronal data.

This is implemented by multiplying the classifier function with smooth constraint terms, which are supported (i.e., attain a nonzero value) only in the empirical range of parameters (see Supplementary Fig. 1). Further, the support of the emerging state indicator function can be confined only to the metric vicinity of the empirical data, or even to the subspace spanned by the data, to keep things conservative. The end result is carving out trajectory space such that only viable trajectories remain – i.e., trajectories that in every form or regard are indistinguishable form the kernel of empirical data the process originated with.

More specifically, say values on the level set manifold are to satisfy the constraint $d_i(g_i(x),\ldots,\theta_i^j,\ldots) < c_i$, where c_i and θ_i^j are empirical estimates of parameters (for example in the case of the Mahalanobis distance $\left\{\theta_i^j\right\}$ would be the set of estimated means and variances of features) pertaining to a given distance function d_i and g_i a transformation (such as a feature transformation). This can be approximated by composing the distance functions with a smooth step function supported in the relevant domain e.g. $\Theta\left(c_i - d_i\left(g_i(x),\ldots,\theta_i^j,\ldots\right)\right)$. Thus:

$$SIF(x) = \Theta\left(c_1 - d_1\left(g_1(x),\ldots,\theta_1^j,\ldots\right)\right)\cdot\ldots\cdot\Theta\left(c_n - d_n\left(g_n(x),\ldots,\theta_n^j,\ldots\right)\right)\cdot Fd(f_1(x)\cdot\ldots\cdot f_k(x))$$

where f_i are feature functions.

Ideally, the parameters used in the constraints should be estimated using points attaining the desired value: $\{x|Fd(x)=c\}$. As the probability of sampling such points is zero, some lenience is called for, that is, deriving the parameters c_i and θ_i^j from a neighborhood of c (e.g., simply taking the points from an experimental condition as a whole).

This procedure is summarized in Supplementary Fig. 1. It is readily apparent that the resulting function would be quite elaborate; one could even go as far a saying monstrous. While this formulation admittedly lacks refinement that future theoretical developments will likely afford, nevertheless great complexity is to be expected. After all, a SIF is to be up to the monumental feat of encapsulating the possible space of trajectories of very complex high dimensional systems.

Appendix B. State indicator functions and reasonable scope, or, what makes for a plausible measure of complexity

As we have argued in detail, systems that give rise to undifferentiated activity spaces are categorically devoid of consciousness. Nevertheless, this is not to say that the structure of a trajectory space alone suffices to realize experience. Rather, only activity trajectory spaces that are parceled by a plausible complexity measure fit the bill. The following somewhat contrived example illustrates this point.

Consider a pseudo-random number generator, which at each clock cycle generates a vector of length n drawn from the uniform distribution on the unit interval. Let there be a mechanism that normalizes these vectors. Further, imagine that the system has an analog dial indicating a positive scalar r by which each such normalized vector is scaled. Given a certain state of the dial, the dynamics will be limited to a sphere, and trajectories of states of length l will lie on the product space of l n-spheres of radius r.

If one were to construe the dial as indicating a complexity measure, its level sets would carve out the activity space of the system into non-trivial state dependant structures (for example the product of 2 *1*-spheres is a torus). Could it be argued based on our analysis that this machine has some sort of proto-consciousness? To appreciate why the structure of trajectory spaces is only as meaningful as a putative complexity measure is, let us imagine that the elements of our system emit fluorescent monochrome light proportional to their level of activation (think transgenic mice). Now, consider two trajectories produced when the dial is set to a level that would enable one to make out the pattern they emit. It just so happens that a movie frame is actually a vector of length n at a given resolution. Thus, if we would take a sequence of frames from, say, Akira Kurosawa's *Rashomon*, we could frame them with a boundary of pixels, and use it to adjust the mean level of luminance to remain invariant across the sequence without changing the relative contrast of the frames. Similarly we could string together a sequence of random frames of equal luminance of the same length. Thus by construction we have two distinct points on the *nl* sphere which have the exact same "complexity" under our measure (Supplementary Fig. 2) – yet by virtually any other measure of complexity fall on the opposite ends of the spectrum (e.g. minimum description length).

Let us now turn to an example that would appear at first to be natural, and is representative of a much larger class of systems: The logistic map, is a function defined on the unit interval, [0, 1] such that $x_{n+1} = rx_n(1-x_n)$.[21] The value of the *control parameter* r determines the behavior of this map when iterated (i.e. $n\to\infty$). For certain values, the map becomes chaotic. Thus, one could argue that a dynamical system that realizes the logistic map has non-negligible consciousness, for values of the control parameter that lead to a chaotic attractor (the set of all values the iterated map attains once n is sufficiently large – which by definition has non-trivial structure), taking the control parameter to be a measure of complexity.

[21] For a discussion of the computational complexity of the symbolic dynamics produced by the logistic map see (Crutchfield, 1994).

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Alas, any plausible complexity measure has to be sensitive to dimensionality both in space (number of elements in the system) and time, for it to be minimally suitable as a complexity measure targeted at experiential state. The reason for this is that a system that gives rise to experience in doing so gives rise to a (virtual) world. The richness of such worlds (that is, of experience) is bounded from above by the dimensionality of activity trajectories both in space and time. In particular, if the dynamics is one-dimensional, the system cannot realize space (to use a ready simile, think of a single pixel). Thus, such a system would trivially have a representational capacity of 0 (which happens to rule out conscious thermostats), as does the logistic map system. Note that this is an inherent limitation, which cannot, moreover, be remedied by wiring together several components whose dynamics obey the logistic map: either they do not communicate, and hence do not form a system, or they do, and thus ipso facto do not realize dynamics described by the logistic map.

It will be useful to analyze this example further, as it actually violates many of the various conditions we spell out that have to be met before complexity analysis can be meaningful in serving as a probe for consciousness. The relation between the structure of a chaotic attractor and that of the resulting trajectory space is not straightforward. For example, attractors are often defined by the set of points given by $\{x_i, x_{i+1}\}$. This corresponds to the minimal trajectory space for a system realizing the logistic map. If one observes (Supplementary Fig. 3a), the 3-dimensional attractor, for three different values of r (4, 3.9, 3.8), it appears to be topologically trivial (it is simply a curve). However, a more careful inspection reveals that this is not the case – in fact, due to the oscillatory nature of the map (for moderate values of r), the attractors are fractured (Supplementary Fig. 3b).

Accordingly, if one analyzes longer trajectories – for example, those of length 10 – the Betti profiles (which represent the multi-scale homology of the data) indeed show a spike at small scales, while being topologically trivial at larger scales (Supplementary Fig. 3c and d). Thus, due to its discrete nature, the system fails to meet one of the most fundamental prerequisites for realizing experience, namely that for at least some representational states, the state dependant trajectory space be locally Euclidean (apart maybe for a subset of points of measure 0), and comprise a single connected component. The reason for this is that any experience can be infinitesimally different in any respect in which its phenomenal content is differentiated, and moreover any normal experience could be smoothly morphed into any other possible stretch of experience through small changes. Thus, experiential systems must arise from dynamics which are invariant in the semantic content they embody in a metric neighborhood of each activity trajectory (Fekete, 2010). Fractured spaces by definition are not such objects.

Further still, the control parameter r does not enable defining a state indicator function. To illustrate this point, we generated three sets of points using the same three values of r as above (4, 3.9, 3.8). We then used these points to build an optimal 3 ways classifier (a Fisher discriminant). However, even with this optimal supervised method, the resulting classifier succeeds only at the moderate rate of 67%. Thus, even if one were able to build a state indicating function for this system, its level sets could not overlap with those induced by the control parameter.

The conclusion from this is that a putative state indicator (complexity measure) has to be measurable in two senses: it must result in separable level sets for distinct values, and, just as importantly, it must be rooted in measurable changes to the system's parameters. Otherwise, it is a construction that is purely in the eyes of the beholder (in being an extrinsic artificial measure that arbitrarily lumps together activity trajectories) rather than a meaningful construct that the system's dynamics realizes.

Finally, the above system, just as any other closed dynamical system, cannot realize experience (Fekete, 2010; Hotton & Yoshimi, 2010, 2011). Rather, only systems that can interact with their surroundings can. Again, this is an inherent limitation for a system realizing the logistic map: if it would also include input terms, it would no longer realize the logistic map.

For completeness's sake, we recapitulate the points made throughout the discussion. A system's dynamics can realize experience only if its dynamics can be partitioned by a state indicator function (SIF) that satisfies the following conditions:

(1) It (and hence the underlying dynamics) must be mutidimensional both in space and time.
(2) The underlying dynamics cannot be discrete.
(3) The dynamics are open. Accordingly, the analysis of state dependant level sets is only meaningful if performed on situated systems.
(4) At least some of the level sets induced by the SIF are topologically non-trivial, and yet unfractured.
(5) A SIF has to be able to attain null states, i.e., the system must have a specific instantiation in terms of parameters for which the putative complexity measure attains a null value – that is, the level set associated with this parameter setting would be devoid of cluster structure. The reason for this is that a model of experience (i.e., a theory-derived set of equations and measures) must also model lack of consciousness.
(6) At least for certain ranges of the SIF, there must be a coupling between the complexity of activity trajectories and the complexity of the space comprising them.
(7) Finally, for a system to realize a conceptual structure, it must exhibit multiscale organization (Fekete, 2010), and thus it should be amenable to a description by plausible complexity measures on several scales.

In conclusion, it can be seen that for a physical system to realize dynamics that meet such criteria is no small feat. Thus, our account fares quite well in terms of the reasonable scope requirement that computational theories of experience have to meet.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.concog.2011.02.010.

References

Amit, D. (1995). The Hebbian paradigm reintegrated: Local reverberations as internal representations. *Behavioral and Brain Sciences, 18*(4), 617–625.
Arieli, A., Sterkin, A., et al (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science, 273*(5283), 1868–1871.
Azouz, R., & Gray, C. M. (1999). Cellular mechanisms contributing to response variability of cortical neurons in vivo. *Journal of Neuroscience, 19*(6), 2209–2223.
Azouz, R., & Gray, C. M. (2000). Dynamic spike threshold reveals a mechanism for synaptic coincidence detection in cortical neurons in vivo. *Proceedings of the National Academy of Sciences of the United States of America, 97*(14), 8110–8115.
Azouz, R., & Gray, C. M. (2003). Adaptive coincidence detection and dynamic gain control in visual cortical neurons in vivo. *Neuron, 37*(3), 513–523.
Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology. *Perception, 1*(4), 371–394.
Baum, E., & Haussler, D. (1989). What size net gives valid generalization? *Neural Computation, 1*(1), 151–160.
Bendich, P., Edelsbrunner, H., et al (2010). Persistent homology under non-uniform error. *Mathematical Foundations of Computer Science, 2010*, 12–23.
Bickle, J. (2006). Multiple realizability. Stanford Encyclopedia of Philosophy.
Blumer, A., Ehrenfeucht, A., et al (1986). *Classifying learnable geometric concepts with the Vapnik–Chervonenkis dimension*. NY, USA: ACM New York.
Brown, R. (2006). What is a brain state? *Philosophical Psychology, 19*(6), 729–742.
Bubenik, P., & Kim, P. (2006). A statistical approach to persistent homology. Arxiv preprint math/0607634.
Cao, Y., Cai, Z., et al (2007). Quantitative analysis of brain optical images with 2D C0 complexity measure. *Journal of Neuroscience Methods, 159*(1), 181–186.
Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines, 4*(4), 391–402.
Chalmers, D. J. (1995). The puzzle of conscious experience. *Scientific American, 273*(6), 80–86.
Churchland, M., Cunningham, J., et al (2010). Cortical preparatory activity: Representation of movement or first cog in a dynamical machine? *Neuron, 68*(3), 387–400.
Churchland, P., & Sejnowski, T. (1992). *The computational brain*. The MIT Press.
Churchland, M., Yu, B., et al (2007). Techniques for extracting single-trial activity patterns from large-scale neural recordings. *Current Opinion in Neurobiology, 17*(5), 609–618.
Churchland, M., Yu, B., et al (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience, 13*(3), 369–378.
Clark, A. (1985). Qualia and the psychophysiological explanation of color perception. *Synthese, 65*(3), 377–405.
Contreras, D., & Llinas, R. (2001). Voltage-sensitive dye imaging of neocortical spatiotemporal dynamics to afferent activation frequency. *Journal of Neuroscience, 21*(23), 9403–9413.
Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2*, 263–275.
Crutchfield, J. (1994). The calculi of emergence: Computation, dynamics and induction. *Physica D: Nonlinear Phenomena, 75*(1–3), 11–54.
de Silva, V., & Carlsson, G. (2004). Topological estimation using witness complexes. In *Proceedings symposium on point-based graphics* (pp. 157–166).
Del Cul, A., Baillet, S., et al (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology, 5*(10), e260.
Dennett, D. (1988). Quining qualia. Consciousness in modern science.
Dennett, D. (2004). *Freedom evolves*. Penguin Books London.
Edelman, S. (1993). On learning to recognize 3-D objects from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15*(8), 833–837.
Edelman, S. (1999). *Representation and recognition in vision*. The MIT Press.
Edelman, S. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences, 6*(3), 125–131.
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 and Theoretical Artificial Intelligence, 20*(3), 181–196.
Edelman, S. (2009). On what it means to see, and what we can do about it. In S. Dickinson, A. Leonardis, B. Schiele, & M. Tarr (Eds.), *Object categorization: Computer and human vision perspectives*. Cambridge University Press.
Edelsbrunner, H., Letscher, D., et al (2002). Topological persistence and simplification. *Discrete and Computational Geometry, 28*(4), 511–533.
Egan, G. (1995). Permutation city, Millennium.
El Boustani, S., & Destexhe, A. (2009). A master equation formalism for macroscopic modeling of asynchronous irregular activity states. *Neural Computation, 21*(1), 46–100.
Fekete, T. (2010). Representational systems. *Minds and Machines, 20*(1), 69–101.
Fekete, T., Pitowsky, I., et al (2009). Arousal increases the representational capacity of cortical tissue. *Journal of Computational Neuroscience, 27*(2), 211–227.
Fisch, L., Privman, E., et al (2009). Neural "ignition": Enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron, 64*(4), 562–574.
Geffen, M., Broome, B., et al (2009). Neural encoding of rapidly fluctuating odors. *Neuron, 61*(4), 570–586.
Geman, S., Bienenstock, E., et al (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1–58.
Gross, B., Walsh, C., et al (2009). Open-source logic-based automated sleep scoring software using electrophysiological recordings in rats. *Journal of Neuroscience Methods, 184*(1), 10–18.
Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*(03), 377–396.
Hasson, U., Nir, Y., et al (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634.
Hobson, J. A., Pace-Schott, E. F., et al (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences, 23*(6), 793–842. 904–1018; 1083–1121..
Holcman, D., & Tsodyks, M. (2006). The emergence of up and down states in cortical networks. *PLoS Computational Biology, 2*(3), 174–181.
Hotton, S., & Yoshimi, J. (2011). Extending dynamical systems theory to model embodied cognition. *Cognitive Science*. doi:10.1111/j.1551-6709.2010.01151.x.
Hotton, S., & Yoshimi, J. (2010). The dynamics of embodied cognition. *International Journal of Bifurcation and Chaos, 20*(4), 1–30.
Hudson, B., & Miller, G., et al. (2009). Mesh enhanced persistent homology. Computer Science Department: 1122.
Hume, D. (2007(1748)). *An enquiry concerning human understanding and other writings*. Cambridge University Press.
James, W. (1976). *Essays in radical empiricism*. Harvard University Press.
Kenet, T., Bibitchkov, D., et al (2003). Spontaneously emerging cortical representations of visual attributes. *Nature, 425*(6961), 954–956.
Lashley, K. (1923). The behavioristic interpretation of consciousness. *Psychological Review, 30*(4), 237–272.
Leznik, E., Makarenko, V., et al (2002). Electrotonically mediated oscillatory patterns in neuronal ensembles: An in vitro voltage-dependent dye-imaging study in the inferior olive. *Journal of Neuroscience, 22*(7), 2804–2815.
Lindsay, P., & Norman, D. (1972). *Human information processing: An introduction to psychology*. New York: Academic Press.
Mach, E. (1886). *1959 the analysis of sensations*. New York: Dover.
Makarenko, V., Welsh, J., et al (1997). A new approach to the analysis of multidimensional neuronal activity: Markov random fields. *Neural Networks, 10*(5), 785–789.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY, USA: Henry Holt and Co., Inc.
Maudlin, T. (1989). Computation and consciousness. *The Journal of Philosophy, 86*(8), 407–432.
Mazor, O., & Laurent, G. (2005). Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. *Neuron, 48*(4), 661–673.
McDermott, D. (2001). *Mind and mechanism*. The MIT Press.
Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(01), 63–81.
Metzinger, T. (2003). *Being no one*. The MIT Press.
Mongillo, G., Barak, O., et al (2008). Synaptic theory of working memory. *Science, 319*(5869), 1543–1546.
O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences, 22*(01), 127–148.
Putnam, H. (1988). *Representation and reality*. The MIT Press.
Quine, W. (1951). Two dogmas of empiricism. *Philosophical Review, 60*, 20–43.
Rabinowitch, I., & Segev, I. (2008). Two opposing plasticity mechanisms pulling a single synapse. *Trends in Neurosciences, 31*(8), 377–383.
Searle, J. R. (1990). Is the brain a digital computer? *Proceedings and Addresses of the American Philosophical Association, 64*(November), 21–37.
Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317.
Smart, J. (2007). The identity theory of mind. Stanford Encyclopedia of Philosophy.
Spivey, J. (2007). *The continuity of mind*. USA: Oxford University Press.
Steriade, M., Timofeev, I., et al (2001). Natural waking and sleep states: A view from inside neocortical neurons. *Journal of Neurophysiology, 85*(5), 1969–1985.
Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, 42.
Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215*(3), 216–242.
Truccolo, W., Hochberg, L. R., et al (2010). Collective dynamics in human and monkey sensorimotor cortex: Predicting single neuron spikes. *Nature Neuroscience, 13*(1), 105–111.
Tye, M. (2007). Qualia. Stanford Encyclopedia of Philosophy.
Vapnik, V. (1995). *The nature of statistical learning theory*. NY: Springer.
Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications, 16*, 264–280.
Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.
Yu, B., Cunningham, J., et al (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Journal of Neurophysiology, 102*(1), 614–635.

Please cite this article in press as: Fekete, T., & Edelman, S. Towards a computational theory of experience. *Consciousness and Cognition* (2011), doi:10.1016/j.concog.2011.02.010