I am what I am *

Shimon Edelman† Tomer Fekete‡

January 7, 2013

Abstract

A central, non-negotiable constraint on any account of phenomenal awareness or experience is that it must be *intrinsic* to the experiencer, rather than depending on an outside interpretation of its structure or function. When applied to computational accounts of phenomenal awareness that are based on dynamical systems theory, this constraint raises a serious conceptual challenge, namely, the need for an intrinsic definition of the "system" in question. In this extended abstract, we discuss some of the issues arising from this challenge.

1 Fundamental constraints on theories of phenomenal experience

Phenomenality — the quintessence of first-person experience — cannot be a matter of attribution or ascription: it must be intrinsic to the experiencer. The distinction between attributed and intrinsic properties of informational systems has been drawn in the past. For instance, in their tutorial introduction to the Information Integration Theory of qualia, Balduzzi and Tononi (2009, p.2) write: "From the perspective of an external observer, the camera chip has a large repertoire of states. From an intrinsic perspective, however, the sensor chip can be considered as a collection of one million photodiodes with a repertoire of two states each, rather than as a single integrated system with a repertoire of $2^{1,000,000}$ states." A forceful statement of the principle of intrinsic phenomenality has been offered by Fekete and Edelman (2011, p.811):

> "[...] Activity [the dynamics of the experiential system] 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."

The insistence on phenomenality being intrinsic is what leads one to seek explanations rooted in dynamical systems theory — a move that can also be motivated on independent grounds, both empirical and philosophical (see, e.g., Skarda and Freeman, 1987; Freeman, 2007 and the contributions in Edelman, Fekete, and Zach, 2012). As Fekete and Edelman (2011) discuss at some length, identifying phenomenality with the dynamics of the experiencer (rather than, say, with its instantaneous states; cf. Smart, 2004), in addition to solving a host of conceptual problems, makes phenomenality intrinsic. Simply put, by computing its future given its past and any external influences, a dynamical system by definition serves as its own interpreter. To do that, it needs time, which is why phenomenality is not a state but a process (Edelman and Fekete, 2012).

*Extended version of an abstract submitted to the 17th Conference of the Association for Scientific Study of Consciousness, 2013.

†To whom correspondence should be addressed. Department of Psychology, Cornell University, Ithaca, NY 14853-7601, USA.

‡Department of Biomedical Engineering, Stony Brook University, Stony Brook, NY 11794-5281, USA.

2 The open questions

Identifying phenomenal experience with the dynamics of the experiencer gives rise to several related issues, as detailed below.

2.1 How to define the system that embodies the experience?

Spivey (2006, p.305) offers the following vivid formulation of the dynamical approach to phenomenality:

> "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."

This view of phenomenality, which is close to the one we recently defended (Fekete and Edelman, 2011), begs the question: "The trajectory of what system?" A moment's reflection suggests that the seemingly obvious answer — "The brain" — is glaringly incomplete.

First, this answer usually tacitly assumes that it's (the dynamics of) the spiking activity of the brain's neurons that matters, without explaining why (the dynamics of) graded potentials, cell metabolism, chemical diffusion, quark-gluon interactions inside the atomic nuclei, and any number of other ongoing processes that are happening inside a living brain should be excluded (cf. Fekete and Edelman, 2012b). This is the question of *level* on which the (intrinsically) relevant dynamics is supposed to be occurring. Assuming that the level of spiking neurons is the only one that counts does not resolve this issue.

We remark that although the question of level arises in computational accounts of any aspect of cognition, in the accounts of phenomenality it is particularly pressing. With regard to any cognitive function, such as visually guided behavior (say), the scientist, acting as an external observer, can rule out the functional involvement of various levels, for instance by showing that information contained in spike trains but not in graded potentials makes a difference for the function under consideration. In comparison, in addressing the brain basis of phenomenal experience, one doesn't have such luxury: what matters here is not whether the inclusion of graded potentials in the model of the brain makes a difference for an external observer, but rather whether or not it makes an intrinsic difference — a point that seems very hard to address empirically.

Second, given that minds are now widely construed as incorporating computations carried out by the physics of the bodies that support them, as well as extending beyond individual bodies (e.g., Thompson and Varela, 2001; Anderson, Richardson, and Chemero, 2012), it is far from clear what, if anything, demarcates the experiencer from the rest of the world. (As a particularly nagging example of a consideration stemming from radical embodiment, one may consider here the effects of gut microbiota on cognition; Cryan and Dinan, 2012.) This is the question of the *boundary* around the system whose dynamics is at stake. Merely assuming a weak coupling between the brain and the rest of the world (Hotton and Yoshimi, 2010) doesn't quite suffice here.

2.2 What about the subsystems formed by subsets of the experiencer's components?

Let us assume for the moment that the issue of demarcating the experiencer is somehow resolved. That would still leave us with the need to explain why the dynamics of the entire N-element (say) system that comprises the experiencer, but not any of its 2^N subsystems, is what the experience is to be identified with. We may call this the *subsystem* question.

Some of the proposed theories of phenomenality (e.g., Tononi, 2008; Fekete and Edelman, 2011) recognize this issue. In particular, Tononi's Information Integration Theory is built on the notion of an information complex: a system whose joint state space contains, by virtue of element interactions, information that cannot be found in the respective state spaces of its elements. A "maximal complex" is then defined in the straightforward fashion, and qualia are attributed to it, but not to its inferiors. (Balduzzi and Tononi, 2009, p.5) write

> "[. . .] Only a complex can be properly considered to form a single entity and thus to generate integrated information."

Unfortunately, as with the first issue above, this formulation begs the question: "Considered by whom?" While Tononi is clearly aware of the need for an intrinsically valid formulation of such claims (e.g., Hoel, Albantakis, and Tononi (2012)), the singling out of the maximal complex is done by fiat and appears to be less than fully convincing (for some philosophical arguments against it, see Schwitzgebel, 2012).

2.3 How to define "the" computation implemented by the experiencer?

The third issue, which is somewhat different from the first two, stems from the multiple realizability of computations. A common feature of the dynamical systems theories of consciousness that we referred to so far is that they identify the explanandum with the computation performed by the experiencer. But what, exactly, is "the" computation here? As Fekete and Edelman (2011, pp.810-811) put it,

> "Given that a 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?"

For an in-depth analysis of this issue, see (Fekete and Edelman, 2012b).

3 Is there a way out?

To recap, what we need to come up with is an intrinsic justification for each of the following explanatory moves:

1. The choice of an implementational *level* or levels on which the account of phenomenal experience resides;
2. The choice of a *boundary* between the experiencer and the rest of the world;
3. The decision to exclude all but one of the many *subsystems* contained in the experiencer system (perhaps the maximal one) from participating in the account.

Can the level, boundary, and subsystem issues be resolved without giving up what seems to us the right framework for explaining phenomenality (and other aspects of the mind), namely, computation in dynamical systems?

3.1 The boundary and the subsystem issues

The latter two questions — whether an intrinsically valid boundary can be drawn around the experiencer and whether the parts or subsystems, if any, of the experiencer have phenomenality of their own — are closely related to each other. Let us take them up first.

3.1.1 Being one with all

One way out of the boundary predicament is suggested by the following passage from (Gier and Kjellberg, 2004), who write in the context of a discussion of the problem of freedom of will:

> "If we cannot call the karmic web free since it lacks a self, by the same token we cannot call it determined, since nothing outside of it is causing it. To the extent that people identify a self, that self is determined by causes outside of it. The more cultivated they become on the Buddhist model, the less they think this way. The less who thinks this way? A question that the European philosopher might ask. Nagarjuna's answer is no one, really. The non-personal web of causes and conditions sheds the delusion, or, rather, ceases to give rise to it. [. . .] Thus while we would assume that there has to be a self in order for there to be freedom, Nagarjuna would say that there is freedom only to the extent that there is not a self."

Mutatis mutandis, one can state with regard to the problem of delimiting the experiencer (subject) of a given experience (the subject's representation of some object) that to the extent that one insists on drawing a distinction between the two — that is, drawing a boundary between the subject and the rest of the universe — the resulting limits would necessarily leave out some of the subject's components. In other words, any definition of the experiencer that encompasses less than the totality of the universe necessarily leaves out certain aspects both of the experiencer and of the experience.

It is interesting to observe that a holistic stance, albeit not as extreme as the one just stated, has been advanced in the context of the issue of the unity of consciousness by Bayne and Chalmers (2003). The philosophical treatments of the unity of consciousness (see (Brook and Raymont, 2012) for a review), which does relate to issues that we raise in the present paper, traditionally tend not to concern themselves with implementational details, let alone with the question of how the unification can arise intrinsically. For instance, (Shoemaker, 1996, p.177) writes: "The visual experience of a spatially extended thing is a synthesis of visual experiences of parts of that thing" — a claim that begs the question "Synthesis by what means?" Partly in response to such concerns, Bayne and Chalmers (2003) propose that phenomenality is holistic to begin with, so that its splitting into aspects, not its integration, needs to be justified, if at all.

3.1.2 Relativistic considerations and an invariance principle

Although the extreme holistic stance on phenomenality does resolve the boundary issue, it amounts to giving up on the challenges it poses, instead of meeting them. Moreover, a closer look at this stance reveals a complication. As far as anybody knows, information in this universe cannot propagate faster than the speed of light in vacuum, a fact whose relevance to theories of consciousness has been pointed out by Edelman and Fekete (2012). Thus, for the purpose of defining "me" as an experiential system, one may ask: should everything that's within my light cone count as part of me? The complication arises from the need to fix the apex of "my" light cone, and to do so in an intrinsically justified manner.

To make the best of this situation, we can try to see the problem of localization of experience as amounting to a constraint on theories of phenomenality. This constraint may be stated as a kind of *invariance principle*, according to which any definition of the phenomenality must be such that the qualia experienced by a system turn out to be the same, no matter where they are localized (i.e., where the apex of the light cone associated with a "qualia probe" happens to be).

The extreme holistic stance satisfies the invariance principle trivially: the entire universe is intrinsically and invariantly conscious. Are there nontrivial solutions to this conundrum? Because the real underlying

problem — finding a natural (intrinsic) way of defining a "system" — is much broader than the corner of science occupied by theories of consciousness, it makes sense to approach it gradually through less loaded examples.

3.1.3 System and subsystem questions in physics

For instance, in simple mechanics we may take inspiration from Galileo's famous thought experiment and think of a candidate system S consisting of a large mass M, connected via an elastic cord to a smaller mass m. How massive would S appear to itself? Note that this question cannot be answered unless time is allowed to roll and S begins to move. It then depends on various factors, including the elasticity constant and the initial tautness of the cord, the direction of the motion of M relative to m, and the time scale of interest (or, equivalently, the length of the cord relative to M's speed). If the cord has initially some slack, then when M begins to move, S initially "feels" M-massive. After a while, it "feels" $(M + m)$-massive or, more generally, $f(M, m)$-massive for some possibly nontrivial $f(\cdot, \cdot)$. An electrical analogy using a circuit that includes a diode (which conducts current in one direction only) readily suggests itself.

It would seem that the two-mass example prompts us to ask, with regard to the elements of a system that jointly give rise to experience by communicating with each other, "Who knows what when?" We must, however, resist such questions, which, like Malach's (2012a) focus on "neural consensus," only leads to more trouble, both computationally (insofar as distributed consensus is deeply problematic; Edelman and Fekete, 2012) and conceptually. When M is moving, but m isn't (yet), does S move? Questions that belong on an aggregate level just cannot be asked of the aggregate's components and still make sense.

This observation suggests that our problems may be resolved through a shift from a detailed to an aggregate or even statistical view of phenomenal experience. Thus, a more apt physical analogy to consider may be a spatially distributed property, in a system consisting not of two elements but of many, as in the case of the Fermi energy level of charge carriers (electrons and holes) in a solid medium — an insulator, a semiconductor, or a conductor. It makes no sense to ask what the Fermi level is for a given atom in the medium (just like it makes no sense to ask what the temperature of an atom is). And what if one asks, "How large is the region of influence that determines the Fermi level at a given test point?" The answer would then depend relativistically (in the sense introduced earlier) on the timing of the propagation of various electromagnetic disturbances through the medium.

3.1.4 I am what I am

The lesson from the preceding discussion is that we shouldn't try too hard to pinpoint and localize processes that just aren't local in any relevant sense, either in space or in time. With regard to the boundary issue, given a spatiotemporal region with the right kind of dynamics (as spelled out in Fekete and Edelman, 2011), we may say that experience happens then and there, without insisting that none of it "spills over" to the rest of spacetime. If the time window is extended and a perturbation is delivered from the outside to this region of space, the dynamics of the stuff that's contained in it will be modified in some manner, and with it the experience. In other words, the ongoing phenomenality of a chunk of brain is what it is; if a stimulus impinges on it from the outside, it will change, with the change spreading out gradually through field propagation and spike trains, eventually becoming something else (Edelman and Fekete, 2012). The boundary issue seems, thus, not to be an issue after all.

Keeping in mind the aggregate nature of phenomenality, as suggested by the physical analogy, we now contend that the subsystem issue is likewise devoid of substance. The solutions proposed for it — such as the maximal complex idea of Tononi (2008) and the similar approach espoused by us in Fekete and Edelman

(2011), which both hold that no subset of some uniquely "largest" set of elements comprising a system can have experiences — seem rather arbitrary in that they privilege one subset over all others. What if we set this notion aside?

Trivially, each of the 2^N subsets of the N-element system must be "doing its thing" for the dynamics of the entire system to be what it is. Translated into the language of experience, this means that for me to feel as I do now, all the subsets of the neurons in my brain must be "humming" just so. Note that the same view holds true also of other aggregate-functional phenomena, such as dance or music. Thus, when I dance, so does my right arm, including all its bones, joints, muscles, tendons, etc. (although the dance of the arm is not as complex as the dance of the entire body, and despite the fact that it would be unable do dance on its own). Likewise, when a choir sings Mahler's 8th symphony, the soprano section — whose make-up may differ, depending on how many singers in this category the producers managed to recruit for this particular performance — sings too (even though its song is not as complex as the symphony of which it is part).

We should not let our intuition, which suggests phenomenal unity and indivisibility, thwart this move to do away with the subsystem issue. Indeed, given that *this* is what it is like to be me now, and given that all my neurons (and all their coalitions, combinations, etc.) are together shaping the dynamics of my brain, for all I know *this* is what it is like to be a chorus of N neurons with just *this* dynamics — and, again trivially, a chorus of 2^N neural cliques, not one of which, not even the maximal one, is in any way a priori privileged.

Importantly, the preceding claim does not imply that my self includes N selves each of which differs from the "whole" by having one neuron less (etc., etc., for other subsets). What saves us from this exponential proliferation of dubious entities is the fact that those sub-selves are an arbitrary imposition from the outside, and thus intrinsically as good as nonexistent. While connected to the rest of the brain, each of the N neurons in question keep contributing to the whole's dynamics, and if disconnected, they would leave a (presumably minimally) changed whole.

3.2 The level issue

The Fermi energy analogy may also help us draw a line in the sand with regard to the issue of the level(s) on which dynamics that makes up a system's phenomenal experience may reside. The qualitative behavior of charge carriers in a solid such as iron or silicon, as determined collectively by the Fermi energy, is effectively insulated from subatomic (nucleus-level) processes (as evidenced by the enormous energy needed to split the stable nuclei of iron or of silicon). In other words, quarks and gluons may cavort ad infinitum inside the nuclei of a solid without having the slightest effect *on any aspect of its dynamics at the level of the nucleus or above*. Thus, given (i) that all evidence suggests that experience *is* affected by levels from the atom upwards, and (ii) that those levels are insulated from the subatomic levels, we may conclude that the latter are irrelevant to experience.

This conclusion still leaves us with a lot of possibilities regarding the relevant dynamics. While great care must be exercised in considering what it means, within this domain, for two systems to be equivalent with regard to "the" computation they carry out (cf. Fekete and Edelman, 2012b), the Fermi energy analogy does offer some hope that certain equivalences could become apparent (the alternative — that no two systems that differ in their physical make-up can ever have commensurable experiences — seems to us counterintuitive). Because it is possible for disparate materials to have equivalent aggregate electrical and thermodynamic properties, Fermi band structure is multiply realizable. Likewise, if experience indeed arises from a certain type of computation, namely, from discernment that takes the form of intrinsic structure of aggregate trajectory dynamics (Fekete and Edelman, 2011), then it too may well be multiply realizable.

References

Anderson, M. L., M. J. Richardson, and A. Chemero (2012). Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science 4*, 717–730.

Balduzzi, D. and G. Tononi (2009). Qualia: the geometry of integrated information. *PLoS Computational Biology 5*(8), 1–24.

Bayne, T. and D. Chalmers (2003). What is the unity of consciousness? In A. Cleeremans (Ed.), *The Unity of Consciousness: Binding, Integration and Dissociation*, pp. 23–58. Oxford: Oxford University Press.

Brook, A. and P. Raymont (2012). The unity of consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012 ed.).

Cryan, J. F. and T. G. Dinan (2012). Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nature Reviews Neuroscience 13*, 701–712.

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

Edelman, S., T. Fekete, and N. Zach (Eds.) (2012). *Being in Time: Dynamical Models of Phenomenal Experience*. Amsterdam: John Benjamins.

Fekete, T. and S. Edelman (2011). Towards a computational theory of experience. *Consciousness and Cognition 20*, 807–827.

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

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

Freeman, W. J. (2007). Indirect biological measures of consciousness from field studies of brains as dynamical systems. *Neural Networks 20*, 1021–1031.

Gier, N. and P. K. Kjellberg (2004). Buddhism and the freedom of the will. In J. K. Campbell, D. Shier, and M. O'Rourke (Eds.), *Freedom and Determinism: Topics in Contemporary Philosophy*, pp. 277–304. Cambridge, MA: MIT Press.

Hoel, E. P., L. Albantakis, and G. Tononi (2012). The 'neural code' from the intrinsic perspective: Quantifying causal power at different spatio-temporal scales.

Hotton, S. and J. Yoshimi (2010). Extending dynamical systems theory to model embodied cognition. *Cognitive Science 35*, 444–479.

Schwitzgebel, E. (2012). If materialism is true, the United States is probably conscious. Unpublished ms.

Shoemaker, S. (1996). *The first-person perspective and other essays*. Cambridge, UK: Cambridge University Press.

Skarda, C. and W. J. Freeman (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences 10*, 161–195.

Smart, J. J. C. (2004). Consciousness and awareness. *Journal of Consciousness Studies 11*, 41–50.

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

Thompson, E. and F. Varela (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences 5*, 418–425.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull. 215*, 216–242.