ACAUSALITY AND THE MACHIAN MIND

John W. Jameson

ABSTRACT: In this paper we propose a mechanism in the brain for supporting consciousness. We leave open the question of the origin of consciousness itself, although an acausal origin is suggested since it should mesh with the proposed quasi-acausal network dynamics. In particular, we propose simply that fixed-point attractors, such as exemplified by the simple deterministic Hopfield network, correspond to conscious moments. In a sort of dual to Tononi's Integrated Information Theory, we suggest that the "main experience" corresponds to a dominant fixed point that incorporates sub-networks that span the brain and maximizes "relatedness." The dynamics around the dominant fixed point correspond in some parts of the system to associative memory dynamics, and to more binding constraint satisfaction dynamics in other areas. Since the memories that we are familiar with appear to have a conscious origin, it makes sense that a conscious moment itself corresponds in effect to what amounts to memory recollection. Furthermore, since Hopfield-like networks are generative, a conscious moment can in effect be seen as a living, partially predicted memory. Another primary motivation for this approach is that alternative states can be naturally sensed, or contrasted, at the fixed points.

KEYWORDS: Brain; consciousness; Tonini's Integrated Information Theory

1 INTRODUCTION

Imagine you are alone in the universe. There is no one, there are no other things, nothing around you but empty space. Now imagine you are spinning. You feel your arms being tugged outward from the centrifugal force. Or will you? How can you say you are spinning if there is nothing else to refer to or interact with? Such imaginings were the inspiration of Mach's Principle: that the inertia of our bodies must be due to all the other stuff out there in our universe.[1]

[1] Mach's Principle was one of Einstein's inspirations for his general theory of relativity and is still seriously regarded, although controversially, by fundamental physicists.

Similarly, is the heft of my thoughts, of the concepts in my (conscious) mind, dependent simply on all the other ones in my mind? Is my mind a closed off universe on its own? This is the assumption of the Machian Mind. The mind is a universe created by neurons and synapses, of interacting electrical charges. It is true that the outside world impinges on my senses, but these senses are still just part of my brain.

Both the present approach based on *relatedness* and Tononi's Integrated Information Theory (IIT) [3] are Machian in their inspiration. However, there is a serious thorn in the side of this inspiration: that qualia seem to have an absolute character to them. That, e.g., the sense of the color green does not, e.g., seem relative to other concepts in the mind. We simply posit that the internal, relative, concepts of the Machian mind provide the proper conditions for *more* (kinds of) qualia.

As an engineer I tend to think of the brain as a machine. But can a machine produce consciousness? If so, this would imply that consciousness itself follows a course completely dependent on the machinery, rendering it but an epiphenomenon. But why would nature bother with something that is essentially impotent? For this reason I think there is something more to consciousness than can be explained by machinery, at least at the level of description we currently have. For example, it may be an emergent property of certain complex systems [18, 20, 44], or it may entail a quantum mechanics [5, 10, 13, 33, 34].

In this paper we explore the possibility that there is at least a great deal of machinery (at least of the type we are familiar with) *supporting and interfacing* with consciousness.[2] It is based fundamentally on the notion that consciousness is broken up into a sequence of *conscious moments* . In this vein we consider the following criteria for the *Consciousness Support Mechanism (CSM)*:

1. Can the mechanism manifest itself over a short time period of a *conscious moment* ?

2. Can pieces of it (sub-CSM's) be put together to form a larger whole, i.e., are they constitutive?

3. Is there some evidence from neuroscience and cognitive science to back it up?

The foregoing is aimed primarily at the avid connectionist with an interest in philosophy of mind (or perhaps a philosopher of mind with an interest in connectionism). The connectionist models are given as simple examples to capture the basic ideas, hopefully with some biological plausibility, at least in spirit.

[2] It should be pointed out, however, that relatedness might be useful for governing the behavioral dynamics of the network system as well as its self-organization, regardless of its status with respect to consciousness, although this may alter some of the constraints placed on the dynamical system (see Sec. 8).

2 INTEGRATED INFORMATION THEORY

Giulio Tononi developed the Integrated Information Theory (IIT) about a decade ago as a measure of the *degree of consciousness* for an information processing system, with details worked out for a neural network model [4]. His theory posits that the degree of consciousness for any set of elements in the network (called a complex) is proportional to the amount of (Shannon) information the complex generates as a whole relative to the amount it would generate if it were not as fully connected in some sense (the latter is called the "minimal information partition"). He calls this relative information *Integrated Information* and designates it with the symbol

Φ. The complex in the system with the highest **Φ** he says "underlies the dominant experience."

Christof Koch [19], a long time champion of Tononi's theory, pointed out that IIT implies that any (causal) information flow *is*, or *entails* , consciousness, which amounts to a version of panpsychism. Searle recently criticized this approach [31, 38] because panpsychism doesn't seem to account for the boundaries for the kinds of consciousnesses we know about: I am here and you are there and our consciousnesses don't overlap. According to IIT, however, the consciousness I am familiar with (as opposed to e.g., my subconscious) corresponds to an endless series of main complexes, which greatly mitigates the possibility of multiple sloppy, overlapping, consciousnesses. Nevertheless, I don't believe IIT accounts for consciousness more for another reason: it implies that consciousness is an epiphenomenon, and thus has no independent ontological status. That neither hurts nor helps from an evolutionary standpoint, but I find it very hard to believe that such a miracle could arise from an indifferent process.

Having said that I think **Φ** may have a lot to do with the *potential* for consciousness. In this case we can see that if integrated information flow is not consciousness itself, then it is possibly the *content* of consciousness. In IIT on the other hand, integrated information is simultaneously the subject and the object, suggesting in effect "I am what I sense and nothing else," with no accounting for a will or agency.

If information itself isn't consciousness as IIT implies, can IIT account for the CSM, i.e., the cognitive mechanism most directly underlying consciousness? There are two issues with this: 1) how would the system know what the "natural" minimum information partition (MIP) is, and 2) how would the system effectively cut the connections corresponding to the MIP in order for the generated information for the partition to be compared to the unpartitioned "actual" network?[3]. It seems very challenging for this to be accomplished biologically, much less in a conscious moment.

[3] Tononi actually uses injected noise as a substitute for a severed connection.

In any case, because IIT doesn't seem useful as a potential CSM, and because it relies on causal information flow, why not consider an opposing kind of dynamics? *In this paper we explore acausal dynamics for the CSM.*

3 CAUSAL VERSUS QUASI-ACAUSAL NETWORK DYNAMICS

We can break down interactions in the neural network of the brain into two broad categories: *causal* and *quasi-acausal* interactions. *Causal* here is defined in the usual sense: something happens here which then creates an effect there and so forth, in which the happenings are typically neuronal activations. A *quasi-acausal* interaction on the other hand refers to an interaction whose outcome, or average result over some period of time, is (almost) independent of the sequential order of its constituent interactions.[4] Acausal dynamics express *structure.*

As an example we consider a neural network consisting of a set of binary nodes connected to each other *bidirectionally*, as shown on the left side of Fig. 1. Such a network is called a *Boltzmann Machine* if the activation of each node is *stochastic* (probabilistic). When the system settles down to its equilibrium dynamics, which may take a long time, the probability of going from one (total system) state to another is the same as going in the opposite direction.[5] Because of the long time spans required for this (directional) acausality to manifest itself (even assuming the system has reached equilibrium), however, this is an implausible CSM according to the first CSM criterion.

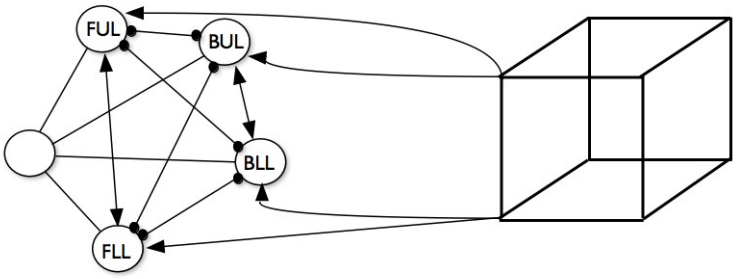

Figure 1*: A Hopfield network with bi-directional connections (left) and the Necker Cube (right) it represents. FUL = front-upper-left, BLL=back-lower-left, etc. Only a few nodes of the network are shown. The connections with arrows are excitory and the ones with dots are inhibitory.*

[4] This is not to be confused with the definition acausal systems used signal processing, e.g., where future signal values are used to calculate a the output of a filter.

[5] This property of graphical models like the Boltzmann Machine is called "detailed balance."

4 THE HOPFIELD NETWORK

If instead the activation of each node in the network in Fig. 1 is *deterministic* instead of stochastic, the network will fairly rapidly settle to a *stable fixed point* , i.e., to fixed activation configuration. The network in Fig. 1 in this instance is a Hopfield Network [16]. Near the fixed point the dynamics are essentially acausal (i.e., quasi-acausal).

The idea behind the fixed point is easy to see mathematically (and these will be the only equations in this paper). Let x represent the state (vector) of the network, where each element of the vector corresponds to a node state (e.g., 1 or 0). If x_t if the state of the network at time t, then we can express the state of the network at time $t + 1$ as:

$$x_{t+1} = R(x_t) \quad (1)$$

where $R(\cdot)$ is a vector function which we call the *relation function* . A fixed point x^* is reached when:

$$x^* = x_{t+1} = R(x_t) \qquad (2)$$

If the fixed point is *stable*, or *attractive*, then small perturbations about x^* still lead the state to x^*. We can consider the state x to be discrete of approximately continuous, as expressed for example through the current firing rate [17].

A Hopfield network can act as an autoassociative memory network. A set of patterns are "imprinted" onto the network through Hebbian learning, and by activating a portion of the same pattern later, the full original pattern will emerge after a short number of cycles.

Hopfield networks can also solve *constraint satisfaction* problems. A classic example is the Necker Cube, shown in Fig. 1, for which there are two ways of perceiving the three-dimensional structure from the image. Each way corresponds to a fixed point of the Hopfield net [27].

However, there are major problems with the standard Hopfield network. As an associative memory it does not have a very large capacity. If too many patterns are added it is easy for spurious patterns to be recalled. As a constraint satisfaction problem, in addition to the chance of settling on a poor (locally optimal) solution, there needs to be a mechanism for labeling the network in the first place. For example, in Fig. 1 the nodes need to be ascribed to the proper abstract entities, and the abstract entities themselves need to be established by the system somehow (see Sec. 6; also [40]). Nevertheless, I think it is likely these issues can be surmounted by more sophisticated network systems (including Hopfield-like networks working with other kinds of networks).

Note that a Hopfield network with stochastic units is essentially a Boltzmann Machine. Conversely, a Boltzmann Machine becomes a Hopfield network when the "temperature" of the former is reduced to zero. Boltzmann machines can learn powerful hidden, or latent, representations which could, e.g., provide the abstract kind of labels needed for the deterministic constraint satisfaction presently postulated as a requirement for a consciousness support mechanism [28]. Because of this natural connection between Boltzmann and Hopfield networks it is natural to link the two in a cognitive/conscious process. However, there are still serious issues with Boltzmann machines. E.g., it usually takes thousands of samples before truly representative samples from the full distribution are obtained (one sample is typically very close to the next sample). This is covered more in Sec. 6.

In any case, I refer to the deterministic dynamics corresponding to conscious moments as arising from *Hopfield-like* networks in this paper. Again, these same networks could operate largely stochastically as a preparation for their deterministic stage for the conscious moment, while also working alongside other networks systems to also set the stage for these moments.

That fixed point dynamics correspond to "perceptual states" is hardly a new idea. The Necker Cube demonstration in fact was motivated by this very fact (and this was done in the 1980's). Recent work by Braun and Mattias [6] parallel some of the larger points presented in this paper. Cao et al. [7] address suggest that global perceptual states correspond to *absorbing states* that are essentially fixed points for a system that are reached through a stochastic (Ehrenfest) process. In Sec. 8.6 we suggest a more deterministic mechanism for this process, mainly because stochastic models take far too many cycles to generate a representative mixture of samples, at least for the models we are familiar with.

In any case, if conscious moments correspond *only* to fixed points, this puts serious constraints on the nature of the fixed points, as we shall discuss. So even if we don't know the underlying nature of consciousness, investigations like the present one might offer insights into the kinds of machinery we should be seeking for supporting the mind.

5 TO KNOW A THING

When my perception, or perspective, of the Necker Cube changes it is disorienting. But when I lock onto to one of the two perspectives its hard to break away from it. My brain seems to have a serious "cling to thing" syndrome. Furthermore, it seems that this is the same phenomenon I experience when I suddenly "know" something in general (even though it is often wrong).

It is presently posited that what we think of as knowing something is the sensing of a stable fixed point. Knowing is the surety of the stable fixed point.

What do I mean by sensing a fixed point? That numerous perturbations of the system about the fixed point, whether stochastic or deterministic, serve to elicit the information or meaning of the fixed point state by contrasting it with (local) alternative states. To quote Dretske [8]:

> ". . . for me to see that the soup is boiling—to know, by seeing, that it is boiling—is for the soup to be boiling, for me to see the soup, for the conditions under which I see the soup to be such that it would not look the way it were not boiling, and for me to believe that the soup is boiling on that basis."

Furthermore, this contrasting of alternative states needs to be done in the present conscious moment. This is the blessing of the attractive fixed point. Such contrasting can be done rapidly and reliably because the deviations are small and the dynamics usually bring it back toward the fixed point.[6] As to the "meaning" elicited by the wiggling around the fixed point: for the Machian mind in any case, meaning can only exist with respect to the rest of the network, through its interaction with it.[7] This is discussed in more detail in Section 77.

This relates to the criticism mentioned earlier regarding IIT: that while (Shannon) information accounts for alternative possible states, there isn't a biologically plausible way to compute them, much less compute them in a conscious moment. IIT gets around this by claiming that consciousness simply is information, and thus doesn't need to be computed.

6 ABSTRACT ENTITIES, BINDING, AND ASSOCIATE MEMORIES

Knowing a thing implies there is a thing to be known, and without things you can't have relations between things, i.e., you can't have *relatedness* (see Sec. 7). However, to the degree that Hopfield-like nets cannot construct (latent) abstractions we need another system for the learning of (hidden) abstractions. Once abstractions begin emerging (or nucleating) the relational Hopfield-like system can help refine their definition and "focus" them on their antecedents.

[6] This is effectively sampling the derivatives of the energy function.

[7] Note also that I have not mentioned meaning as it related to the *value* of states from a reinforcement learning perspective (the value of a state is the essentially expected future reward for that state) . But if it is possible to ascribe a value to every possible states, then the sensing of alternative states as suggested here would also mean sensing the corresponding alternative (reward) values.

Consider the Necker Cube again. The diagram in Fig. 1 presupposes basic concepts of space such as front, up, and left, as well as low level features for lines and corners, and relations between them. These concepts need to be discovered and represented in a distributed yet segregated manner before the system starts relating them. Even more, we need mechanisms for labeling the lower level embodiments of entities like edges to their corresponding high level concept representations, i.e., we need to *bind* the higher level abstraction with their lower level antecedents. To do this we need a lot of other machinery.

6.1 Abstractions

The field of machine learning and neural networks has been largely dedicated to developing methods for discovering latent or hidden representations of data. But in typical models it's hard to pin down what the "things" *are* in the hidden representations because they are mixed up together in a distributed fashion on the same set of nodes (or neurons).[8]

One recent example of progress in this area is that by Le Roux et al. [26], whose system of Boltzmann machines can delineate things (classes) in the data. In particular, their system effectively *binds* abstractions to their antecedents in images according to their textural characteristics. A somewhat more elaborate version of this system might be able bind the pixels belonging to a car in the image to the model for the car in the network system (effectively segmenting the car in the image).

Another example is the HDP-RBM system of Tenenbaum and Salaktudinov [29], where Boltzmann machines are combined with a hierarchical Bayesian model with Dirichlet priors that can discover classes and superclasses, i.e., abstract things, in the data.

6.2 Binding

We expect that the types of relations admissible for Eq. 1 are of the binding type. A good current review for the binding problem is presented by Feldman [9]. For example, *feature binding* entails binding elements of an object, such as geometric and color properties of parts of an object to the object representation itself [39]. The approach of Le Roux et al. [26] (mentioned in the previous section) falls under this category although would probably not qualify as working as part of the CSM since they use stochastic networks.

[8] This is the partly the motivation of sparse, over-complete representations. [22], i.e., to unmix the hidden representation as much as possible while offering some modicum of coarse coding.

For more sophisticated brains there is also the variable/role filler kind of binding [32, 41, 43]. If we desire network connections (i.e., portions of R in Eq. 1) to represent more general relations, e.g., "is the uncle of," the number of possible relations between possible combinations of objects grows astronomically, too large even for machine as complex as the brain.[9] Thus, if such relations as "is the uncle of " are to be included it appears they must be in the form of neural activities[10] rather than connections between neurons, with the roles and fillers carried out by the binding relations (connections).

6.3 Associative memories, living memories, loops within loops.

As mentioned previously, a Hopfield net can serve as an associative memory. Each stored pattern corresponds to an attractive fixed point. Implementing the "constraints" in this case is trivial: the weights between the nodes of the network are updated via Hebbian learning for each stored pattern. The *recollection* of a store pattern entails the convergence to an attractive fixed point. In this sense, ascribing a conscious moment to a fixed point, to the degree that it corresponds to memorization and not a binding constraint (although one could look at a memory in this case as a pattern bound to itself), implies that a conscious moment corresponds to a kind of "living memory."

I think this is a compelling argument for the present hypothesis that conscious moments correspond to fixed points: *since all the memories that we are familiar with stem from conscious moments. What better way to ensure that the conscious moment is memorized effectively than to have itself correspond to a kind of (living) memory?*

Note that consciousness seems to entail a kind of short term memory trace, a kind of gestalt across a short time window. Interestingly, this might imply a kind of recursion though, a trace of a trace (of a trace of a trace...) if you will. It puts a new spin on Hofstadter's idea that "I am strange loop" [15].[11]

7 RELATEDNESS AND THE MACHIAN MIND

Instead of Integrated Information (Φ) as a measure of the degree of consciousness, we propose a measure called *Relatedness* ($\mathcal{R}$). The "dominant experience" in terms of relatedness is expected to correspond to a large complex (similar in spirit to that proposed by IIT), which in turn is expected to consist of myriad sub-complexes. As

[9] An example of this approach is linear relational embedding [23], where relations and objects are embedded in vector/matrix spaces. This approach has a connectionist flavor to it in that objects and relations are represented in distributed fashion and learned from data.

[10] Probably via coarse coding rather than lone grandmother cells.

[11] The loop in Hofstadter's book refers to the model of the self, which contains a model of the self, which contains a model of the self, etc.

discussed earlier, both Φ and $\mathcal{R}$ implicitly suggest that the existence of mind is a result of mutual interactions of all the elements within the brain, which is very much the flavor of Mach's Principle.

7.1 Defining Relatedness

How can we define $\mathcal{R}$ mathematically? At this early stage we have not committed to a definition, nor in fact may this actually be necessary. Nevertheless I present, qualitatively, one approach based closely on the analogy with the Mach's Principle: suppose that object/abstraction representation can be accorded some amount of "mass," and that higher level abstractions have more mass than lower ones because they are bound to more other concepts than lower level ones (on average). We also expect that the more mass a concept has, the less rapidly its status changes (on average). Indeed, this is the inspiration behind *Slow Feature Analysis*, a form of unsupervised learning for temporally sparse, abstract representations [11, 45].

Imagine we have a set of representations that are bound via a fixed point, as shown in Fig. 2, where lower levels representations are at the bottom. Also suppose that for each fixed point, or conscious moment, all the bound representations are accorded a fixed increment in mass (thus we must suppose that the masses of all representations slowly decay in the absence of stimulation). Note that since the binding connections are bidirectional there is no sense of hierarchy, i.e., all the representations (R's) bound by straight lines in the figure achieve the same boost in mass for the bound state shown. However, *over time* higher level concepts will be bound to a more lower

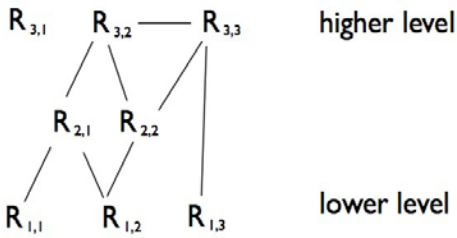

Figure 2: *Bound representations.*

level ones than vice versa. This is accounted for by increasing the *intrinsic mass* of the representation (thus making it slower).[12]

[12] One way of making a representation slower would be to simply have a longer time constant for integrating its input, although there are probably better ways.

This is indeed very Machian: the intrinsic mass of an object (representation) is achieved to the degree it interacts with everything else. As in slow feature analysis, the system is forced from the trivial solution of infinite mass everywhere by forcing subsytems to respond to the local input (e.g., sensory input at the lowest level). Also keep in mind that the Hopfield-like networks are expected to be used in conjunction with other kinds of networks for generating many of the representations accessed by the former (see Sec. 6.1).

This is a very preliminary idea for expressing relatedness mathematically. Ideally, whatever the measure, it could be used to guide the cognitive/mental dynamics as well as the self-organization (learning) of the network system.[13]

7.2 Qualia and relatedness

Contrary to the Machian viewpoint, qualia have an absolute character to them: the quality of red does not seem relative to other concepts in my mind. On the other hand, the fact that I am seeing this particular red thing now, and only I am seeing it, is more in line with the Machian paradigm. Let it suffice for now that it seems that both the relative character of meaning ascribed to (most) internal representations on one hand, and the absolute character of qualia on the other, matter in subjective experience. Related abstractions can simply add more qualia to mix. For example, when I have a thought, I "feel" the thought.[14]

In any case, representations closely tied to sensory information should have a special status in terms of how they influence the measure of relatedness, given that they are otherwise at the lowest levels (e.g, in slow feature analysis the network is simply forced to respond to the input). If we think of the concept of the self as being at a very high level, while holding that qualia are quintessentially subjective, then it is rather paradoxical that the low level sensory driven qualia seem to be, at least structurally, very distal from the model of the self.

7.3 Relatedness and Integrated Information

Interestingly, Tononi found that **Φ** is *lowest* for attracting fixed points in Hopfield network dynamics, and low but not quite as low for nearby states [3]. This is not surprising since **Φ** is based on causal information flow and the present relatedness measure is based on a situation that is (almost) acausal. Quoting Tononi from this paper: "the elements are working against each other locally (at the level of couples),

[13] As far as I know, Tononi has been mute on the possible use of **Φ** for self-organization.

[14] Some philosophers call such sensations as these "cooked feels" (and not really qualia) as opposed to "raw feels," which are associated more directly with our senses. In this paper I call both types qualia.

and so the system does not cohere as a single global entity." This is certainly true for his definition of integrated definition, but for the present measure of relatedness these locations are the only ones that mean anything. This situation is settled if we accept that we are 1) either talking about two network systems, or 2) possibly the same system but at different times (operating in different modes).

7.4 The constitutive property

It appears that consciousness has a constitutive nature (item 2 in CSM criteria list): I could, say, add another arm to my body, and if I were to connect the nerves up just right and so forth it would become part of my consciousness. In the present version of relatedness this means *tying the fixed points together, melding them into a single attracting fixed point in a higher-dimensional space.*

8 BUT CONSTRAINT SATISFACTION IS HARD

In general, solving a constraint satisfaction problem on a finite domain is an NPcomplete problem (can't be done basically). However, neural networks like the Hopfield network can find *locally* optimal (possibly very poor) solutions (attracting fixed points) fairly rapidly (see Sec. 4).

Incidentally, the fixed points don't need to be completely re-computed for consecutive conscious moments. One fixed point is usually very close to the previous one and thus relatively very little computation is needed. It's only when the picture (or the perspective as in the case of the Necker Cube) drastically changes that they need to be recomputed, although probably not from the ground up (or should we say from the top down) though since the new picture will likely reappear in the same higher order context in the dominant fixed point.

Constraint satisfaction on a massive scale is an extremely challenging problem to say the least. It may in fact need consciousness to make it work. Nevertheless, there are a number of constraints (on the binding relations) that should be able to greatly expedite this process (this list is sure to grow; also see Sec. 6):

1. During waking periods, the system must sustain a steady stream of attracting fixed points (default mode).
2. In general, enforce of only those constraints that can be enforced (ignoring the rest).
3. The fact that similar things relate (bind) similarly to other things.
4. The fact that the components of similar things typically have a one-to-one mapping (binding) to each other.

5. Enforce local, intra-module constraints much more stronger initially, then slowly strengthen the coupling between the modules (inter-module constraints).
6. The construction of knowledge (of the fixed point kind) from scratch, starting with knowledge of one's own body e.g.
7. Tune the system in order to facilitate *amplitude death* (see Sec. 8.6).
8. Others items from Gestalt Psychology [42], such as laws of proximity, closure, symmetry, common fate, continuity, and good gestalt, might apply but some of these are less obvious in terms of binding constraints.

8.1 Maintain at least a minimal flow of fixed points (default mode)

This is the prime directive . Continuity of consciousness from moment to moment is obviously very strongly enforced in real minds.[15] Thus the system is effectively forced to only relate things that it can (i.e., establish corresponding attracting fixed points), at least for things that will correspond to conscious awareness.

This could relate to the default mode network [25], which holds sway when you're not actively doing a task: what happens when you close your eyes and clear your mind. What "structure" is the dominant fixed point generating? Probably mostly core things such as your body in space and the sense of self.

8.2 Enforce only constraints that can be enforced

This is strongly related to the previous item, but applies much more broadly, not to just the (default) continuity requirement.[16] It supports the mind's tendency to "cling to things." As I mentioned before, an ideal situation is when all the neurons involved in the attracting fixed point will be strongly on or off. For those that are in the middle, not quite sure which way to go, these neurons might need to effectively switch themselves off so as not to participate in the constraint satisfaction process.

8.3 Similar things relate similarly to other things, etc.

If we assume the network consists of arrays of similar modules, e.g., cortical columns in V1, an effective constraint on the (bi-directional) connections (relations) between the

[15] I use "continuity" a little loosely here. It is entirely possible that it is more like a movie film in the sense of having a sequence of moments. But clearly here the prime directive still applies.

[16] This is similar to a feature of Adaptive Resonance Theory that encourages the network to only learn to predict what it can and to ignore the rest (17).

columns would be that they be similar (in architecture, synaptic efficacy etc.). This can be achieved partially via mechanisms for topographical mapping.

Interestingly, Tononi discovered that coupled information processing networks generate *large* Φ (integrated information) when those modules are similar to each other [37] (p. 221). Perhaps a more prosaic way to express this is "a thing interacts most effectively with things similar to it." An engineer might call this *impedance matching* . In any case, an example of a connection between Φ and $\mathcal{R}$ might be thus stated as: all else being equal, two things that (causally) interact strongly are likely to (acausally) relate similarly to other things.

8.4 The identity relation

This is how the perception of, for example, one automobile can map the (stored) perception of another automobile, e.g., by mapping the hood of one to the hood of another, the front left tire of one to the other, and so forth. These kinds of relations imply analogies and can thus form foundations for generalization (e.g., developing a category such as "automobiles").

Another example is how conscious entities relate to their immediate past versions. Such relations apply to innumerable modalities and subsystems throughout the brain. We can imagine that the drive of the overall dynamics is to maximize relatedness, the fast route is through this kind of association. Could this be why it seems like my consciousness is a kind of short memory trace in time? A gestalt across space but also across a small moving window of time (see also Sec. 6.3)?

Another example, perhaps the "anchor" for the dominant fixed point, is the model of the self. For this we have the simplest (and strongest) of all relations (constraints) possible: that between the self and the version of the self just a moment ago. I call this the *strong identity relation*. This self-identity only has meaning in the face counter-factuals (again, "sensed" by wiggling about the fixed point).

8.5 Building up knowledge: start with the basics

"Knowledge" here refers to the accumulation of memories corresponding to fixed points.[17] How does the system start building knowledge at the beginning? First consider the sense of space and time and the poor, lonely neuron. It does not "know" where it is. Rather, spatial information is "understood" *implicitly* [5], through movement of the body and corresponding sensory experiences [36], and the

[17] "Knowledge" to most connectionists, on the other hand, typically refers to a more passive, latent, form such as synaptic efficiencies, network architecture, and so forth.

construction of relationships of many neurons with other neurons (in this case in various topographical maps e.g.).

8.6 Amplitude death

Amplitude death is an emergent property whereby a system of many coupled oscillators enters a *global attracting fixed point* [1, 2, 24, 30]. The dynamics are emergent in the sense that the isolated or uncoupled systems do not exhibit stationary dynamics. This also works for coupled *chaotic oscillators* (with dissipative coupling), but in this case the process first entails a conversion of chaotic oscillations to periodic or quasi-periodic oscillations [30].

This process starts with synchronous firing of the neurons before homing in on the fixed point (amplitude death). I have not investigated this thoroughly enough yet to reach a conclusion, and indeed there are some technical hurdles, but it is conceivable that amplitude death could provide an additional pressure for global constraint satisfaction (see the next section). This in fact could be a reason for fast gamma oscillations: as a stepping stone to amplitude death, in this case corresponding to a dominant fixed point.

9 INTERACTING WITH A STRICTLY ACAUSAL CONSCIOUSNESS

Tegmark calls the attractive fixed point a "hyper-classical" state [35]; I prefer to call it "quasi-acausal" because my current intuition is that consciousness itself is strictly acausal and that we wish Hopfield-like network dynamics to mesh with it.

If we think of it from a quantum mechanical perspective, we could also say that the cycling through states of the classical system as it approaches a fixed point is analogous to the superposition of all possible states in the quantum system, and the strict acausality corresponding to (non-local) entanglement (of ions) [13] (or phonons).

10 AGENCY, ATTENTION, AND THE QUALE OF THE WILL

As discussed in Sec. 2, if consciousness is not an epiphenomenon, then consciousness—as an entity itself—must have some means to express actions. Such actions serve to guide conscious awareness itself (through shifting attention),[18] as well as to move the organism in order to thrive and survive. If actions are to be expressed by a Hopfield-like network through the CSM, they must be part of the dominant fixed point.

To the degree that the action is generated by another parallel network system (see Sec. 6.1), or by consciousness as an independent entity, the Hopfield-like network

[18] A good example of the former is the conscious shifting of one's perspective of the Necker Cube.

might serve as merely a message transferal mechanism and not contribute any influence on the action selection itself. As such, the incorporation of the action into the fixed point might be accomplished simply by binding the nodes of the action's (very distributed) representation together associatively via fast weights.

On the other hand, to the degree that the action is computed as the result of a (binding) constraint satisfaction process leading to the dominant fixed point, the action can be seen as a kind of prediction (since that is its usual role in perception). Hawkins proposed such a correspondence between prediction and action in his theory of hierarchical temporal memory [14].

As we've mentioned a few times in this paper, the "will" seems to be part and parcel of consciousness. But if both qualia and the will are so fundamental to consciousness, then it is natural to ask: is there a quale associated with the will? When I make a decision about something I may hear my thoughts as words in my head (and I may feel the thoughts in a sense too). These are qualia.[19] But these words were probably generated by my zombie cognition [21]. What made that first choice that got the cognition going in the first place? What *quale* can I associate with that moment? The answer does not seem obvious. If the will is expressing itself in every conscious moment, to experience a feeling with every expression of the will might be overwhelming. Nevertheless, if I had to guess I would say it is close to what it feels the same as what it feels like to be me. Part and parcel indeed.

11 FINAL REMARKS

In this paper I have argued that conscious states are *supported* by the activity around *stable fixed points* of predominantly deterministic dynamics of neural network complexes, driven to maximize relatedness ($\mathcal{R}$). The preference for deterministic as opposed to stochastic dynamics results primarily from the intuition that alternative states need to be expressed in the conscious moment, although this is not so crucial to other points in the paper. The flavors of relatedness include, per the conscious moment, auto-associative memory recall and the binding of concepts and relations to their constituent components. A tentative, qualitative, measure for $\mathcal{R}$ was proposed, inspired by slow feature analysis and the Machian perspective.

The fact that short term memories correspond to conscious moments I believe provides strong evidence for the fixed point hypothesis, in the sense that conscious moments themselves are a kind of currently active (living) memory. Simply attempting to imprint a pattern via Hebbian learning hardly guarantees a successful recall of the

[19] Again, in this paper we regard "raw feelings," or sensory qualia, and "cooked feelings" like the "feeling of a thought" both as qualia (see Sec. 1).

pattern. But if a conscious moment is a kind recollection itself, the reliability of its later recollection later is immensely enhanced.

If conscious moments indeed correspond to fixed points, this might be observable in the brain in the form of a sequence of sustained patterns spanning the brain, each lasting on the order of the conscious moments. Scanning technology will need to be advanced considerably, however, before we could expect to see this.

Jameson Robotics El Cerrito, CA

REFERENCES

[1] Delay is a death sentence. *Phys. Rev. Focus 6* (Oct 2000), 15.

[2] Anastassiou, C. A., Montgomery, S. M., Barahona, M., Buzsa´ki, G., and Koch, C. The Effect of Spatially Inhomogeneous Extracellular Electric Fields on Neurons. *The Journal of Neuroscience 30*, 5 (Feb. 2010), 1925–1936.

[3] Balduzzi, D., and Tononi, G. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Computational Biology 4*, 6 (2008).

[4] Balduzzi, D., and Tononi, G. Qualia: The geometry of integrated information. *PLoS Computational Biology 5*, 8 (2009).

[5] Bohm, D. *Wholeness and the Implicate Order*, reissue ed. Routledge, Nov. 2002.

[6] Braun, J., and Mattia, M. Attractors and noise: twin drivers of decisions and multistability. *NeuroImage 52*, 3 (sep 2010), 740–51.

[7] Cao, R., Braun, J., and Mattia, M. Dynamical features of stimulus integration by interacting cortical columns. In *BMC Neuroscience, Abstracts from the Twenty Second Annual Computational Neuroscience Meeting: CNS*2013* (2013), BioMed Central Ltd, BioMed Central Ltd.

[8] Dretske, F. *Seeing and Knowing*. International library of philosophy and scientific mehtod. University of Chicago Press, 1969.

[9] Feldman, J. The neural binding problem(s). *Cognitive Neurodynamics* (Sept. 2012), 1–11.

[10] Freeman, W. J., and Vitiello, G. Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics. *Physics of Life Reviews 3*, 2 (June 2006), 93–118.

[11] George, D. *How the Brain Might Work: A Hierarchical and Temporal Model for Learning and Recognition*. PhD thesis, Stanford, CA, USA, 2008. AAI3313576.

[12] Grossberg, S. Adaptive Resonance Theory: how a brain learns to consciously attend, learn, and recognize a changing world. *Neural networks : the official journal of the International Neural Network Society 37* (Jan. 2013), 1–47.

[13] Hameroff, S., and Penrose, R. Consciousness in the universe: A review of the Orch OR theory. *Physics of Life Reviews 11* (Mar. 2014), 39–78.

[14] Hawkins, J., and Blakeslee, S. *On Intelligence*, adapted ed. Times Books, Oct. 2004.

[15] Hofstadter, D. *I am a strange loop*. Basic Books, New York, 2007.

[16] Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences 79* (1982), 2554–2558.

[17] Hopfield, J. J. Neurons with Graded Response Have Collective Computational Properties like Those of Two-state Neurons. *Proceedings of the National Academy of Scientists 81* (1984), 3088–3092.

[18] Jones, R. *Analysis & the Fullness of Reality: An Introduction to Reductionism & Emergence*. CreateSpace Independent Publishing Platform, 2013.

[19] Koch, C. *Consciousness : confessions of a romantic reductionist*. MIT Press, Cambridge (Mass.), 2012.

[20] Laughlin, R. *A Different Universe: Reinventing Physics from the Bottom Down*. Basic Books, 2006.

[21] Libet, B. Reflections on the interaction of the mind and brain. *Progress in Neurobiology 78*, 35 (2006), 322 – 326. The Contributions of John Carew Eccles to Contemporary Neuroscience.

[22] Olshausen, B. A., and Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by v1. *Vision Research 37* (1997), 3311–3325.

[23] Paccanaro, A., and Hinton, G. E. Learning hierarchical structures with linear relational embedding. In *NIPS* (2001), T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., MIT Press, pp. 857–864.

[24] Palazzi, M. J., and Cosenza, M. G. Amplitude death in coupled robustchaos oscillators.

[25] Raichle, M. E., and Snyder, A. Z. A default mode of brain function: A brief history of an evolving idea. *NeuroImage 37*, 4 (2007), 1083 – 1090.

[26] Roux, N. L., Heess, N., Shotton, J., and Winn, J. M. Learning a generative model of images by factoring appearance and shape. *Neural Computation 23*, 3 (2011), 593–650.

[27] Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. E. Schemata and sequential thought processes in pdp models. In *Parallel Distributed Processing. Volume 2: Psychological and Biological Models*, J. L. McClelland, D. E. Rumelhart, and PDP Research Group, Eds. MIT Press, Cambridge, MA, 1986, pp. 7–57.

[28] Salakhutdinov, R., and Hinton, G. Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (2009), vol. 5, pp. 448–455.

[29] Salakhutdinov, R., Tenenbaum, J. B., and Torralba, A. Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Mach. Intell. 35*, 8 (2013), 1958–1971.

[30] Saxena, G., Prasad, A., and Ramaswamy, R. Amplitude death: The emergence of stationarity in coupled nonlinear systems. *Physics Reports 521*, 5 (Dec. 2012), 205 – 228.

[31] Searle, J. R. Can information theory explain consciousness? *The New York Review of Books* (January 10, 2013 Issue).

[32] Shastri, L., and Ajjanagadde, V. From simple associations to systematic reasoning: a connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences 16* (1993), 417–494.

[33] Stapp, H. P. Quantum theory and the role of mind in nature.

[34] Stapp, H. P. The quantum-classical and mind-brain linkages: The quantum zeno effect in binocular rivalry.

[35] Tegmark, M. The importance of quantum decoherence in brain processes.
CoRR quant-ph/9907009 (1999).

[36] Terekhov, A. V., and O'Regan, J. K. Space as an invention of biological organisms. *arXiv preprint arXiv:1308.2124* (2013).

[37] Tononi, G. Consciousness as integrated information: a provisional manifesto.
The Biological bulletin 215, 3 (Dec. 2008), 216–42.

[38] Tononi, G., Koch, C., and reply by Searle, J. R. Can a photo-diode be conscious? *The New York Review of Books* (March 7, 2013 Issue).

[39] Treisman, A. Solutions to the binding problem: progress through controversy and convergence. *Neuron 24*, 1 (Sept. 1999).

[40] Uwents, W., Monfardini, G., Blockeel, H., Gori, M., and Scarselli, F. Neural networks for relational learning: An experimental comparison. *Mach. Learn. 82*, 3 (Mar. 2011), 315–349.

[41] Velik, R. From single neuron-firing to consciousness–towards the true solution of the binding problem. *Neurosci Biobehav Rev 34*, 7 (2010), 993–1001.

[42] Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., and von der Heydt, R. A century of gestalt psychology in visual perception: I. perceptual grouping and figure-ground organization. *Psychological bulletin 138*, 6 (Nov. 2012), 1172–1217.

[43] Wendelken, C., and Shastri, L. Multiple instantiation and rule mediation in shruti. *Connect. Sci. 16*, 3 (2004), 211–217.

[44] Werner, G. Consciousness viewed in the framework of brain phase space dynamics, criticality, and the renormalization group.

[45] Wiskott, L., and Sejnowski, T. J. Slow feature analysis: unsupervised learning of invariances. *Neural Comput 14*, 4 (Apr. 2002), 715–770.