Why Scott should stare at a blank wall and reconsider (or, the conscious grid)

Scott Aaronson has taken upon himself to show that integrated information theory (IIT) of consciousness is wrong. As an honorable mention, he concedes that IIT is at least wrong precisely. Because Scott's post is lucid, clearly argued, and constructive in spirit, it calls for a direct response. Moreover, many of the comments on his blog indicate that Scott's challenge should be taken as an opportunity to clarify some aspects of IIT that may not be so easy to grasp. Readers interested in an exposition of IIT should look here, here, and here. A discussion of the neurobiological evidence for IIT is found here.

Scott's argument

To summarize Scott's main argument:

- The best any theory of consciousness can hope for is to say which systems are conscious and which are not (he calls this "the pretty-hard problem")
- At a minimum, such a theory should make predictions that are in line with commonsense intuitions – for example, humans are conscious, some animals probably too, but rocks and crowds probably not
- More specifically, Scott's intuition assures him that certain systems – for example, expander graphs made of simple logic gates – are definitely not conscious. This is because such systems, even very large ones, are quite easy to characterize mathematically and lack computational complexity (though he admits that, due to wiring issues, it is not so easy to actually build large ones)
- IIT defines a quantity - integrated information PHI – that is meant to measure the quantity of consciousness
- Scott then argues that PHI can be very large for large expander graphs and similar systems
- Therefore, IIT is wrong

Is IIT indeed wrong, and has Scott provided a definitive mathematical knock-out? Or did Scott make some mistake in calculating PHI?

A few clarifications before coming to the main point:

- Scott does not consider the latest formulation of IIT and PHI (3.0, here), so his arguments about normalization, for instance, are not relevant (normalization is not necessary in 3.0). Similarly, he does not consider how IIT addresses the problem of identifying the appropriate spatio-temporal grain at which PHI should be calculated (atoms, molecules, neurons, brain areas, people, and so on, see here and here). However, his main point that certain systems that are simple – in the sense that they are easy to describe - could have large values of PHI, still stands. In fact, the numbers for IIT 3.0 would be even "worse"

- Scott considers at some length the Vandermonde matrix. As he himself realizes, this is an abstract mathematical entity, and IIT deals explicitly with actual physical systems, not mathematical idealizations, so Vandermonde is not directly relevant. However, as he says, this cannot rescue IIT, because there are systems that can be implemented physically, such as large expander graphs, which presumably could have high PHI, and that is all it takes to invalidate IIT
- Resorting to expander graphs is actually overkill. This is because systems that are even simpler to describe than expander graphs, for example a 2D lattice of identical logic gates (a "grid") could also achieve very large values of PHI. So things are "even worse" for IIT
- Along the same lines, Scott argues that some systems with high PHI may not only have a structure that is simple to describe, but they may perform computations that are also just as simple to describe, such as parity checks. In fact, the situation for IIT is actually "worse", since it allows for a large 2D grid to be conscious even if it were doing nothing, with all gates switched off at a fixed point. Thus, if IIT can be invalidated by an expander graph doing not much at all, it can be invalidated all the more by a mere grid doing absolutely nothing.

Let us stick to the worst-case scenario, and consider the less forgiving calculation of PHI in IIT 3.0, a large 2D grid of physically implemented logic gates, and assume that all the logic gates are off. Here is what IIT says with respect to Scott's specific challenge:

- Scott's mathematical argument is right: certain systems whose structure and function are easy to describe from the extrinsic perspective of an observer, such as expander graphs performing parity checks, or worse, grids doing absolutely nothing, may in fact have a large value of PHI if they can be built to be large enough (again, they must be actual physical structures)
- Because of their extreme structural and functional "simplicity", they apparently fit Scott's "commonsense" intuition that they cannot possibly be conscious.
- However, Scott's "commonsense" intuition that such simple systems cannot possibly be conscious is wrong and should be revised.

Let me now briefly explain why, and why the specific example of expander graphs or grids provides evidence that is actually in favor of IIT, rather than against it. This can be shown if one considers our own phenomenology, the postulates of IIT that translate fundamental phenomenological properties of consciousness (axioms) into mathematical formulations (postulates) applicable to the physical world, and the resulting explanatory and predictive power with respect to neurobiological facts.

Before proceeding, an important caveat is in order: to understand the significance of what follows, one really needs to understand the premises of IIT – its axioms and postulates, why it is essential to start from phenomenology, why the current mathematical formulation is the way it is, and appreciate the theory's explanatory, predictive, and inferential power. For this reason, even the impatient reader should read at last this and hopefully consider reading this.

Pre-theoretical intuitions can be wrong or contradictory

As some of the blog's commentators have already remarked, it can be dangerous to rely too much on one's pre-theoretical intuitions, however strong they may seem. Examples in science are numerous, starting with the strong intuitions people once had that the earth must be still and the sun must revolve around it, or that the earth cannot be round because otherwise we would fall off. Concerning consciousness, the reliability of pre-theoretical intuitions is even worse, because different people often hold radically different ones. Most of us agree that, since we ourselves are undeniably conscious, people who are built and behave like us are likely conscious, which is a good start. But from there intuitions diverge. Faced with an unresponsive patient, some people's intuition suggests that the patient is conscious (maybe because the eyes are open, or because they think that a glint of emotion can be detected from the patient's expression), while others are convinced that there is nobody home (witness the controversy over Terry Schiavo). Scott easily grants consciousness at least to certain animals, presumably by analogy, since they are built and behave a bit like us. Others would categorically deny this, based on the intuition that consciousness requires self-consciousness, which may be absent in most animals, not to mention newborn babies. Many people are dead certain that simple things like photodiodes are obviously not conscious, while many others are dead sure that consciousness of some sort pervades the cosmos. Some people are certain that non-biological machines cannot possibly be conscious, while others (functionalists like Dennett) are convinced that, if their behavior is indistinguishable from ours in every respect, down to the micro-functional level, then they must be as much (or as little) conscious as ourselves. Other pre-theoretical intuitions are that consciousness requires a physical body, or the capacity to learn from experience, or abstract thought (Descartes), or great "intelligence", or the ability to "report" internal states, or language, or emotion, or the capacity to recognize humor, or ongoing interactions with the environment, or with a social network. Or, for that matter, take Scott's typical computer scientist's intuition: things that are computationally simple (or that may look impressive but can be shown to be simple, either in their structure or in what they do) cannot possibly be conscious.

Why one needs a theory

Since intuitions about who or what may be conscious vary wildly among different people, it seems hard to rely on them to make progress. A better strategy is to try and develop a proper theory of consciousness, one that considers not dubious intuitions, but, *first*, *facts about our own consciousness and, second, facts about the underlying brain mechanisms*.

Phenomenological axioms

A proper theory of consciousness, according to IIT, must consider first and foremost the essential properties of the phenomenon that needs to be explained - experience itself, our own - which is the only one we know directly. IIT identifies five such essential properties (*axioms* of consciousness). Briefly: 1. Each experience exists intrinsically, independent of external observers (existence); 2. It is composed

of many aspects (composition); 3. It is the specific way it is, differing in its own way from countless others (information); 4. It is unified, because it cannot be decomposed into independent components (integration); 5. It is singular, because there is no superposition of multiple experiences with more or less content, flowing at faster or lower speeds. Crucially, each of us can be quite certain of the essential properties of his own experience – I know directly what it is like to be me. By contrast, no one can be really trust his intuition about what it might be like to be some other entity, like a bat, an octopus, or an expander graph. Making meaningful inferences in such cases requires a theory – mere intuitions are not to be trusted.

Postulates about physical mechanisms

According to IIT, the axioms of consciousness must then be translated into *postulates* about how the physical world should be organized to support the essential properties of experience. Briefly: 1. A system of mechanisms exists intrinsically if it can make a difference to itself, by affecting the probability of its past and future states, i.e. it has causal power (existence); 2. It is composed of submechanisms each with their own causal power (composition); 3. It generates a conceptual structure that is the specific way it is, as specified by each mechanism's concept – this is how each mechanism affects the probability of the system's past and future states (information); 4. The conceptual structure is unified - it cannot be decomposed into independent components (integration); 5. The conceptual structure is singular – there can be no superposition of multiple conceptual structures over the same mechanisms and intervals of time. The proposed postulates should be as parsimonious and coherent as possible, and should be given a mathematical form. In the end, IIT can be summarized by the following statement: an experience is *identical* with a maximally irreducible conceptual structure or quale: the quale completely specifies both its quality (the set of concepts in the quale is the content of consciousness) and its quantity (the value of irreducibility Φ^{max} of the quale is the level of consciousness). According to IIT, the quantity and quality of an experience are an *intrinsic* property of a complex of mechanisms in a state – the property of shaping the space of possibilities (past and future states) in a particular way, just as it is intrinsic to a mass to bend space-time around it.

Note the sequence here – one starts from the axioms of phenomenology, and derives postulates about possible physical mechanisms. This is the exact opposite of what is usually done: consider physical mechanisms, whether brains or expander graphs, and wonder how they could possibly generate experience. As David Chalmers has argued convincingly, the problem of going from the brain to experience is indeed so hard that it may actually be impossible to solve. On the other hand, argues IIT, it may be hard, but not impossible to go from experience to the brain (or to expander graphs).

Explanatory power

Next, a theory's postulates must be able to explain, in a principled and parsimonious way, at least those many facts about consciousness and the brain that are reasonably well established and non-controversial. For example, we know that our own consciousness depends on certain brain structures (the cortex) and not

others (the cerebellum), that it vanishes during certain periods of sleep (dreamless sleep) and reappears during others (dreams), that it vanishes during certain epileptic seizures, and so on. Clearly, a theory of consciousness must be able to provide an adequate account for such seemingly disparate but largely uncontroversial facts. Such empirical facts, and not intuitions, should be its primary test.

Predictive power

If the theory does a good job at such facts, one can also use it to make novel predictions that can also be tested empirically. For example, IIT predicts that only manipulations that interfere with the ability of the cortex to integrate information should lead to unconsciousness (e.g. certain general anesthetics such as propofol), while manipulations that affect the brain heavily but do not interfere much with information integration should leave us conscious (e.g., ketamine at certain doses). Or it predicts that if the corpus callosum that connect the two hemispheres were progressively inactivated, there would be a discontinuity at which a single consciousness would split into two independent ones. Despite the inevitable technical difficulties in studying consciousness, the theory will either be corroborated or falsified depending on how its predictions fare, as with any scientific theory.

Inferential power

Assuming IIT continues to be alive and well, it would then become, as is also common in science, our *best inferential tool* to make predictions about situations that are hard to judge using conventional criteria and are therefore highly controversial. These include, for example, patients that are behaviorally unresponsive but show residual cortical activity, newborn babies, animals with alien behaviors and very different brains from ours.

The importance of counterintuitive predictions

As is also common in science, some of the most stringent tests of a theory are predictions that are highly counterintuitive. Note that such predictions must derive directly from the heart of IIT (its axioms and postulates) – they must be consequences, not premises of the theory. In other words, one cannot simply augment the postulates with ad hoc prescriptions that are not grounded in phenomenology, but merely accommodate some cherished intuition (ensuring, say, that expander graphs, or the Unites States of America are not conscious by design).

Here are a few examples of counterintuitive predictions of IIT:

- A photodiode (with memory) is minimally conscious.
- Systems that may seem extremely complicated by most standards, including biological ones, may be unconscious if they do not have the right architecture. For example, a highly interconnected network may nevertheless produce only local maxima of PHI, rather than a single global maximum

- Systems that have a purely feed-forward architecture are unconscious, even though they may be functionally equivalent to systems that perform the same functions consciously
- A conventional computer running a program that behaves just like a human being (think of the movie "Her") would be unconscious, even though it may provide "reports" indistinguishable from a conscious human (Scott's intuition seems to agree on this, but many others disagree)
- Even a conventional computer running a virtual simulation of Scott's own brain down to all its neurons and synapses would be unconscious (and of course it would "report" the same things Scott would)
- On the other hand, a neuromorphic designed in a certain way could be as conscious, or more conscious, than a human being
- A functional cerebral cortex that happens to be silent – no neuron is sending any signal (spike) to any other neuron – would nevertheless be conscious.
- Even a simple but large 2D grid could be highly conscious (this was discussed here, here, and here). Worse still, it could be conscious even if it were doing "nothing" (all its gates off), and even if it were disconnected from the rest of the world (no external inputs and outputs).

Luckily, some of these counterintuitive predictions are in principle testable. Moreover, in some cases we already have some suggestive evidence. One example is the cerebellum, which has 69 billion neurons or so – more than four times the 16 billions of the cerebral cortex - and is as complicated a piece of biological machinery as any. Though we do not understand exactly how it works (perhaps even less than we understand the cerebral cortex), its connectivity definitely suggests that the cerebellum is ill suited to information integration, since it lacks lateral connections among its basic modules. And indeed, though the cerebellum is heavily connected to the cerebral cortex, removing it hardly affects our consciousness, whereas removing the cortex eliminates it.

Let us now turn to another example of a counterintuitive prediction for which we may already have some suggestive evidence, namely that certain kinds of 2D grids may be conscious.

Why certain grids may be conscious

What follows is a summary of why, starting from IIT rather than from some pre-theoretical intuition, it seems quite possible that even seemingly "simple" (easy to describe, i.e. low Kolmogorov complexity) physical structures such as grids may have experience. The argument has four points, which here will be discussed briefly and without illustrations or proper discussion of the literature (details in a forthcoming paper on IIT and the phenomenology of space). To illustrate the advantage of a theoretically motivated approach, rather than intuition, we will consider: 1. some aspects of phenomenology; 2. some of the implications of IIT's postulates; 3. explanatory power; 4. predictive power.

1. Phenomenology: 2D space is a very large, highly structured part of our phenomenology - much of what we experience, we experience in space

It is now time to do what the title exhorts Scott to do - stare at the blank wall, or better still, at a large featureless screen. If we succeed in doing so for a few seconds, unencumbered by extraneous objects or thoughts (some meditators say they are very good at this), we should approximate the pure experience of 2D space, in this case specifically visual 2D space. Two important observations are in order:

First, when we stare at the blank screen for few seconds we are actually quite conscious of the featureless 2D space it encompasses. Somebody could argue that if the screen were to show instead a painting full of colors and objects we would be even more conscious. That may well be, but clearly the difference in the level of consciousness between seeing a blank screen and a painting must be small compared to the difference between experiencing the blank screen and falling unconscious to the ground. It is also worth pointing out that not only are we conscious of the featureless screen even though it contains nothing interesting and nothing is happening on it, but we can be conscious of it even if there is no screen at all, for example if we happen to dream of it – no input/output from/to the external world needed.

Second, if one thinks a bit about it, the experience of empty 2D visual space is not at all empty, but contains a remarkable amount of structure. In fact, when we stare at the blank screen, quite a lot is immediately available to us without any effort whatsoever. Thus, we are aware of all the possible locations in space ("points"): the various locations are right "there", in front of us. We are aware of their relative positions: a point may be left or right of another, above or below, and so on, for every position, without us having to order them. And we are aware of the relative distances among points: quite clearly, two points may be close or far, and this is the case for every position. Because we are aware of all of this immediately, without any need to calculate anything, and quite regularly, since 2D space pervades most of our experiences, we tend to take for granted the vast set of relationship that make up 2D space.

And yet, says IIT, given that our experience of the blank screen definitely *exists*, and it is precisely the way it is - it *is* 2D visual space, with all its relational properties - there must be physical mechanisms that specify such phenomenological relationships through their causal power.

2. Some implications of IIT's postulates: A large 2D grid properly built can give rise to a large conceptual structure of high PHI that specifies, through its causal mechanisms, at least some of the relationships that characterize 2D space

Let us now consider a large 2D grid and apply to it the postulates of IIT (as prescribed in the original publications (here). Importantly, in what follows we are going to make use of what the postulates of IIT imply not just about the *quantity of experience* (or level of consciousness, which is expressed by a single number *PHI*), but also about its *quality* (or content of consciousness, which is expressed by a *maximally irreducible conceptual structure* or quale (a "shape" made of concepts in

concept space). Remember that the heart of IIT is the claim that *an experience is a maximally integrated conceptual structure* (here).[1]

For the sake of the argument, then, let us assume that the 2D grid is made of actual physical elements (say, 1000x1000) that are intrinsically capable of being in at least two different states (for example integrate-and-fire neuron-like units that can be on or off), that they have appropriate input-output functions (for example fire if the majority of the inputs are on), that they have appropriate interconnections (for example each neuron is connected in a lattice to its near neighbors through excitatory connections), and that there is enough time for the relevant causal interactions to occur (say a few msec). Let us also assume that all the elements of the grid happen to be off. Finally, let us also assume that the grid has no inputs or outputs to the outside world (sort of a "dreaming" grid).

Over a certain time interval, each element of the grid specifies its *cause-effect repertoire* – how that element's mechanism being in its current state (off) affects the probability of possible past and future states of the grid itself. If the cause-effect repertoire is maximally irreducible (as indicated by phi^{max}, which measures the irreducibility of individual concepts), it constitutes the element's *concept*, a first-order concept. In essence, such a concept might specify, for example, that the element's near neighbors were likely off a moment in the past and will likely be off a moment in the future. Since there are 10^6 elements, the grid specifies 10^6 first-order concepts.

Moreover, any two nearby elements sharing inputs and outputs might specify a second-order concept, because their joint cause-effect repertoire can be shown to be irreducible to that of the individual elements (as measured by phi>0). In essence, such a second-order concept links the probability of past and future states of the elements' neighbors. By contrast, no second-order concept is specified by elements that don't share inputs and outputs, since their joint cause-effect repertoire are reducible to that of the individual elements. Similarly, subsets of three elements, as long as they share input and outputs will specify third-order concepts linking more of their neighbors together, and so for subsets of four elements, and so on.

In summary, analyzing the grid along the postulates of IIT shows that a properly constructed 2D grid can generate a *conceptual structure* made of a large number of elementary and higher order concepts organized in a particular way. Moreover, this conceptual structure is highly *irreducible* (as indicated by high PHI^{max}, which measures the irreducibility of an entire conceptual structure). This is because, no

[1] It is often forgotten that IIT has considered both quantity and quality of consciousness from the start (here). As Chalmers correctly points out in the comments, a theory of consciousness must account not just for whether a physical system is conscious and how much, but for which particular experience it generates, and what makes it what it is and not something else. To pick an obvious example, when we watch a movie, our brain generates throughout more or less the same quantity of consciousness (PHI), yet it specifies a different experience for every different frame.

matter how one partitions the grid, many concepts would be destroyed (PHI measures the distance between the conceptual structure of the intact and partitioned system), just as implied by Scott's argument for expander graphs. Also, the grid can specify this kind of irreducible conceptual structure even if all its elements are off ("doing nothing"). Finally, it does so even if it has no inputs/outputs from/to the external world and is therefore not "processing" any information or doing anything "intelligent" – it is just specifying what it is like to be such a grid in the space of its possible states (which in IIT is called concept space).

Finally, what matters for the present purposes is that an analysis of the conceptual structure specified by such a 2D grid would in essence reveal the following: i) the conceptual structure *identifies uniquely* the specific causal role of each element of the grid, based on its causal relationships to other elements; ii) it *orders* the elements in two dimensions, thereby specifying their relative position, based on the specific subset of the other elements with which it forms higher-order concepts; iii) it specifies the relative *distance* among any two elements, which is expressed by the lowest order concept they both contribute to.

While this is just a sketch of what the analysis would reveal, it should at least suggest that the conceptual structure specified by such a 2D grid would be nicely suited to specify *experienced 2D space* which, as discussed in 1., *consists of distinct spatial locations, their relative ordering, their distances, and so on, all of which are immediately given in our consciousness*.[2] One may also see that the causal relationships that make up 2D space obtain whether the elements are on or off. And finally, one may see that such a 2D grid is necessary not so much to *represent* space from the extrinsic perspective of an observer, but to *create* it, from its own *intrinsic* perspective.[3]

3. Explanatory power: Many cortical areas are arranged like grids, and the mapping of features on such grids seems closely related to the organization of experience

Decades of neuroanatomy and neurophysiology have revealed that many parts of the cerebral cortex itself are organized as grids. In neuroscience, these are called maps because elements in these maps often bear a clear correspondence with some parameters of the environment that vary in a regular manner. In the simplest case, "topographically" mapped visual cortical areas, such as primary visual cortex (V1),

---

[2] It is interesting to consider the differences in terms of causal power and associated conceptual structure between a grid of intrinsically bistable neurons interacting through fixed anatomical connections such as the one considered here, and other physical substrates, such as a pool of water, a crystal, a gas, and so on. Tegmark (here) gives an interesting assessment of spin glasses.

[3] In fact, it is precisely because we have the experience of visual space from the *intrinsic* perspective – in all its relationships – that makes it easy for us to describe it from the *extrinsic* perspective: one just needs two Cartesian axes X and Y, with some natural numbers mapped on them, say 1 to 1000, and then one can calculate relative position, distances, and so on, with simple algorithms. But this is only because we can presuppose the conscious observer (ourselves).

map Cartesian X-Y positions in visual space using an approximate log conformal mapping. That is, the map is not a strict isomorphic rendition of XY space, for example the fovea is usually much more densely represented than the periphery of the visual field. Also, neurons in such maps do not stand for "pixels" of XY space, but for some specific feature that is mapped topographically, for example the orientation of edges. Many higher-order visual areas, such as V2, V4, V5, V8, and various parietal areas, are also organized like maps, though typically at much lower resolution that V1. Also, areas dealing with non-visual modalities, such as touch or sound, are also organized like maps – for example, primary auditory cortex maps the frequency of sound in an orderly way. Importantly, cortical 2D maps are not just maps, but grids – that is, each element in the map is linked by lateral connections to other elements of the map. Typically, these connections are arranged in a topographic manner, meaning that elements that are closer to each other are more strongly connected, and elements that are far away may not be connected at all. In very rough terms, then, the connectivity of many cortical maps resembles that of a 2D grid with near neighbor connections.

Crucially, for many of these cortical grids, decades of investigations have shown that the way features are mapped on each grid bears a striking resemblance with the way phenomenological distinctions are organized within experience. To pick the simplest example, *where* activations occur on various visual cortical areas (say, V1/2 and V5/MT) can predict *where* certain features are experienced in subjective 2D space. And this is true not only when we are looking at a screen, but also when we imagine or even dream of things. Countless experiments show this to be the case not only in the visual modality, but also in touch, audition, and so on. So there is a clear correspondence between position on 2D cortical grids and position in 2D phenomenological space. In fact this psycho-physical correspondence is by now so obvious that we do not even pay much attention it, without wondering much what the underlying explanation might be. As indicated in 2., IIT can in principle provide exactly this kind of explanation, as long as one understands that ultimately the correspondence is not with cortical space and patterns of activity, but with the associated conceptual structures.

To be sure, there are also many cortical areas that most likely contribute to experience although they are not organized in any obvious topographic manner. Many are areas that care about categorical concepts, such as faces, objects, and so on. Neurons in a face area respond in a position-invariant manner, that is, independent of its spatial location and size. Clearly, categorical concepts such as faces, not to mention concepts that are even more abstract, contribute a great deal to the richness of consciousness. Nevertheless, within consciousness such categorical concepts are usually referred to a particular location in space (a face is always experienced at a particular location and size). Thus, it is likely that the binding between spatial and categorical attributes of an object requires feed-forward and feed-back connections between areas extracting spatial invariants and areas subserving spatially organized features.

4. Predictive power: Topographically organized areas of the cortex are a large contributor to consciousness, and lesioning or stimulating such areas affects experience directly.

For many of the grid-like areas mentioned above, there are a large number of experiments suggesting that they contribute directly to experience, as can be established by lesion and stimulation. Depending on which cortical map is damaged, specific deficits follow, and usually experience itself is altered in characteristic ways. For example, a lesion of primary visual cortex (V1) makes people blind.[4] Conversely, electrically stimulating V1 produces visual "phosphenes" whose perceived spatial location depends on the anatomical site of stimulation. The simplest interpretation of such data is that V1 is essential for generating fine-grained aspects of our experience of visual 2D space. An alternative interpretation, which has led to many experimental tests, is that V1 is only needed indirectly as an input to higher areas, which are then directly responsible for conscious vision (e.g. here). Whatever the final verdict, these higher visual areas are also organized topographically, and we have every reason to believe that, as regards the neural correlates of consciousness (NCC, e.g. here), the buck actually stops with some grid-like area. For example, through lesion and stimulation experiments, it appears that area V5/MT may be directly responsible for our perception of movement in visual space. Similarly, lesions in other topographically organized areas severely impair the overall perception of 2D space (Balint syndrome). In sum, while the study of the neural correlates of consciousness is fraught with experimental and interpretive difficulties, it seems clear that several topographically organized cortical maps likely contribute directly to the quality of our experience, and that manipulating such maps alters our consciousness.[5]

Before concluding, two additional remarks are in order. First, as we have seen, the way to test IIT is not against intuitions, but against the facts of phenomenology and

---

[4] even though, at least in monkeys and some so-called *blind-sight patients*, residual visual behavior remains in the absence of any professed visual experience.

[5] It is an intriguing possibility, apparently made already by Kolmogorov and Barzdin, that cortical connectivity at large may actually resemble an expander graph. Note also that lateral connections can be organized differently, more or less grid-like, in different cortical areas. Moreover, even within each area, different layers of cells are interconnected differently. For example, the cortical "grid" appears to be much denser in superficial than in deeper layers. Finally, it is worth pointing out that topographically mapped structures outside the cerebral cortex, primarily the superior colliculus, also show a grid-like connectivity. It is often assumed that such structures may at most contribute to unconscious vision, but the intriguing possibility should also be considered that they may be conscious in their own right, though not as much and only of 2D space, should also be considered (cf. here). After all, split brain patients demonstrate that the dominant (speaking) consciousness can coexist with a non-dominant one within the same skull, but is not aware of it (here). In IIT a secondary consciousness within the same brain is called a "minor complex".

their empirical substrate. And additional, decisive tests are certainly possible. For example, one could in principle add non-neural hardware with the proper interconnectivity and double the extent of a grid-like visual area that is thought to contribute to the experience of 2D space. If such an addition were to increase the associated conceptual structure and PHI without a corresponding expansion of experienced 2D space and associated increase in consciousness, then IIT would be in serious trouble (granted, this is not an easy experiment to perform, but it is not out of the question).

Second, it is of course possible that, despite the overwhelming evidence, a direct contribution to different features of experienced 2D space by grid-like cortical areas may have nothing to do with their being grid-like, but with some other property. Or that their-grid-like structure may be necessary, but not sufficient to ensure their contribution to experience, and that some additional property may be required, thereby ensuring, say, that grids within the brain would be conscious, but not so grids outside of the brain. Here one can either pick one's favorite candidate for such special property, say a particular firing frequency, a particular kind of synapse, some remarkable computational feature, and so on. The burden, however, is to show that one's favorite candidate can account for the phenomenology of consciousness, including the particular structure of phenomenal 2D space, that it has high explanatory and predictive power, and that it does so better than IIT. Otherwise, adding one's favorite feature is ad hoc and unparsimonious.

Or one can claim that we will never know, because even after we were to account for every aspect of experience in an empirically sound and consistent manner, we would still be left with the hard problem: explaining how the brain can give rise to consciousness. In that case, remember what was emphasized at the beginning and bears re-emphasizing: IIT does not start from brains or grids, but from the phenomenology of experience itself, and *then* asks what it would take for physical systems to account for its properties. Without *presupposing* our own consciousness and its very real properties, including the phenomenology of 2D space, we will never squeeze consciousness out of a brain, a grid, or an expander graph. To recapitulate then, the argument for conscious grids is as follows:

1. Phenomenologically, 2D space has an extremely rich relational structure: scores of distinct spatial locations, their relative ordering, their distances, and so on. This relational structure is repetitive and thus easy to describe extrinsically, but it exists nonetheless.
2. According to IIT, a large 2D grid, properly constructed, can generate a large conceptual structure that specifies through its causal mechanisms many relationships among its elements (their unique identities, relative ordering, distances, and so on). This suggests that it would take the kind of conceptual structure that can be specified by a 2D grid to give rise to the phenomenology of 2D space
3. A large portion of the cerebral cortex is organized like a 2D grid, and the mapping of features on such grids seems closely related to the organization of experience

4. Manipulating these grids can alter or abolish the corresponding aspects of experience, including the overall experience of 2D space.

Hence, if we consider phenomenology and empirical evidence, considering simple grids (and expander graphs) actually supports IIT, rather than invalidating it.

To conclude, whether one likes grids or not, think highly of them or not, and no matter what your intuition tells you about their level of consciousness, you should begin to take them seriously, as it seems that our own experience of 2D space requires a grid to create it. True, even though we can try to approximate it by staring at a blank screen, we do not experience exactly "what it is like to be a single, isolated 2D grid", because we are made of multiple interconnected grids, and of many additional structures that extract categories out of grids. And yet, if we trust a theory that starts from phenomenology and is supported by empirical evidence more than unreliable and unsupported intuitions, our best inference should be that, if a 2D grid is large and well built, it could be quite conscious, though perhaps a bit boring and not that intelligent.[6]

A few final comments

- IIT is an evolving framework: the precise way of measuring PHI, the correct metric for concept space, and even the correct set of axioms and corresponding postulates are still being developed and refined. For example, should there be an independent axiom explicitly about time? Or can time, like space, be derived from the other axioms, based on how perturbations are applied to reveal a system's cause and effect repertoires? Should causal power itself be a phenomenological axiom, or is it implied by the existence axiom, as in the current formulation? Most certainly, the mathematical formulation of IIT will also have to be augmented and improved.
- IIT is all for mathematics, not against it. In his blog, Scott remembers (as best as he could reconstruct it) that when we briefly met in Puerto Rico I would have said: "it's wrong to approach IIT like a mathematician." What I actually said then, and many times before, is that IIT does not start by being infatuated with a particular mathematical quantity – say one that seems to capture an intriguing kind of complexity (PHI or anything else) – and then asking whether that quantity might also capture that intriguing thing that is consciousness. As Scott also remembers (this time correctly), I reminded him that IIT starts from phenomenology itself (a point that cannot be overemphasized), and tries to give a

[6] Because it would lack invariant concepts (disjunctions of conjunctions), which are essential for generalization and survival in a complex environment, and the intrinsic causal relations between all concepts are highly symmetric. Note also that, while IIT predicts that consciousness - i.e. integrated information PHI and the number of concepts within the quale – will usually grow with adaptation to a complex environment (see here, here, Albantakis et al., in preparation) and thus often co-vary with "intelligence" (whatever exactly that might be), one can certainly be highly conscious yet do nothing intelligent (for example, stare at a blank wall).

mathematical expression to the fundamental properties of experience. In short, IIT does not start from mathematics hoping to explain phenomenology, but rather it starts from phenomenology to end with mathematics. Hence mathematical arguments are not only welcome, but are essential to the development and revision of IIT. To provide just one quote: "Ultimately, however, the goal of the present framework is to offer a principled way to begin translating the seemingly ineffable qualitative properties of experience into the language of mathematics (see last sentence here)."

- Finally, I am truly grateful to Scott for having made an effort at challenging IIT in his blog. In this way he has offered a good opportunity for IIT to challenge back some of his own intuitions and, I suspect, those of many others. This of course is how theories are sharpened, intuitions reconsidered, and progress can be made.

I thank Larissa Albantakis, Melanie Boly, Chiara Cirelli, Lice Ghilardi, Jaime Gomez, Erik Hoel, Christof Koch, Will Mayner, Armand Mensen, Masafumi Oizumi, Shuntaro Sasai, and Max Tegmark for their feed-back and suggestions.